

# Document made available under the Patent Cooperation Treaty (PCT)

International application number: PCT/JP2005/017696

International filing date: 27 September 2005 (27.09.2005)

Document type: Certified copy of priority document

Document details: Country/Office: JP  
Number: 2005-131992  
Filing date: 28 April 2005 (28.04.2005)

Date of receipt at the International Bureau: 17 November 2005 (17.11.2005)

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b)



World Intellectual Property Organization (WIPO) - Geneva, Switzerland  
Organisation Mondiale de la Propriété Intellectuelle (OMPI) - Genève, Suisse

日 本 国 特 許 庁  
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日  
Date of Application: 2 0 0 5 年 4 月 2 8 日

出 願 番 号  
Application Number: 特 願 2 0 0 5 - 1 3 1 9 9 2

パリ条約による外国への出願  
に用いる優先権の主張の基礎  
となる出願の国コードと出願  
番号  
J P 2 0 0 5 - 1 3 1 9 9 2  
The country code and number  
of your priority application,  
to be used for filing abroad  
under the Paris Convention, is

出 願 人  
Applicant(s): 松下電器産業株式会社

2 0 0 5 年 1 1 月 2 日

特許庁長官  
Commissioner,  
Japan Patent Office

中 嶋



【書類名】 特許願  
【整理番号】 7048070042  
【提出日】 平成17年 4月28日  
【あて先】 特許庁長官殿  
【国際特許分類】 G06F 17/30  
【発明者】  
    【住所又は居所】 大阪府門真市大字門真 1 0 0 6 番地 松下電器産業株式会社内  
    【氏名】 稲葉 光昭  
【発明者】  
    【住所又は居所】 大阪府門真市大字門真 1 0 0 6 番地 松下電器産業株式会社内  
    【氏名】 菅野 祐司  
【特許出願人】  
    【識別番号】 000005821  
    【氏名又は名称】 松下電器産業株式会社  
【代理人】  
    【識別番号】 100097445  
    【弁理士】  
    【氏名又は名称】 岩橋 文雄  
【選任した代理人】  
    【識別番号】 100103355  
    【弁理士】  
    【氏名又は名称】 坂口 智康  
【選任した代理人】  
    【識別番号】 100109667  
    【弁理士】  
    【氏名又は名称】 内藤 浩樹  
【先の出願に基づく優先権主張】  
    【出願番号】 特願2004-345392  
    【出願日】 平成16年11月30日  
【手数料の表示】  
    【予納台帳番号】 011305  
    【納付金額】 16,000円  
【提出物件の目録】  
    【物件名】 特許請求の範囲 1  
    【物件名】 明細書 1  
    【物件名】 図面 1  
    【物件名】 要約書 1  
    【包括委任状番号】 9809938

## 【書類名】 特許請求の範囲

### 【請求項 1】

構造化文書を管理するデータベース構築装置において、  
構造化文書にユニークな文書番号を割り当てるとともに構造の解析を行う入力文書解析部と、  
前記入力文書解析部の解析結果に基づいて、前記構造化文書に出現する各要素名に対してユニークな要素名 I D を割り当てて要素名辞書に登録する要素名登録部と、  
前記入力文書解析部の解析結果に基づいて、前記構造化文書に出現する各祖先パス名に対してユニークな祖先パス名 I D を割り当てて祖先パス名辞書に登録する祖先パス名登録部と、  
前記入力文書解析部の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名 I D と分岐順の情報を少なくとも含む要素出現情報を、要素名 I D をキーとして要素出現情報格納部に登録し、かつ、着目要素の出現する文書番号と文字位置と要素名 I D と分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名 I D をキーとして祖先パス出現情報格納部に登録する出現情報登録部と、  
を有することを特徴とするデータベース構築装置。

### 【請求項 2】

前記入力文書解析部の解析結果に基づいて、構造化文書に出現する各属性名に対してユニークな属性名 I D を割り当てて属性名辞書に登録する属性名登録部を有し、  
前記出現情報登録部が、前記入力文書解析部の解析結果に基づいて、着目属性の出現する文書番号と文字位置と祖先パス名 I D と要素名 I D と分岐順の情報を少なくとも含む属性出現情報を、属性名 I D をキーとして属性出現情報格納部に登録することを特徴とする請求項 1 に記載のデータベース構築装置。

### 【請求項 3】

前記出現情報登録部が、前記入力文書解析部の解析結果に基づいて、要素実体テキストおよび属性値から切り出された部分文字列に関し、出現する文書番号と文字位置と祖先パス名 I D と要素名 I D と属性名 I D と分岐順の情報を少なくとも含むテキスト出現情報を、切り出された部分文字列をキーとしてテキスト出現情報格納部に登録することを特徴とする請求項 1 に記載のデータベース構築装置。

### 【請求項 4】

前記要素出現情報は、着目要素の出現する文書番号と文字位置と祖先パス名 I D と分岐順と空要素順の情報を少なくとも含み、  
前記祖先パス出現情報は、着目要素の出現する文書番号と文字位置と要素名 I D と分岐順と空要素順の情報を少なくとも含むことを特徴とする請求項 1 に記載のデータベース構築装置。

### 【請求項 5】

前記要素出現情報は、着目要素の出現する文書番号と文字位置と祖先パス名 I D と分岐順と空要素順の情報を少なくとも含み、  
前記祖先パス出現情報は、着目要素の出現する文書番号と文字位置と要素名 I D と分岐順と空要素順の情報を少なくとも含み、  
前記属性出現情報は、着目属性の出現する文書番号と文字位置と祖先パス名 I D と要素名 I D と分岐順と空要素順の情報を少なくとも含むことを特徴とする請求項 2 に記載のデータベース構築装置。

### 【請求項 6】

前記要素出現情報は、着目要素の出現する文書番号と文字位置と祖先パス名 I D と分岐順と空要素順の情報を少なくとも含み、  
前記祖先パス出現情報は、着目要素の出現する文書番号と文字位置と要素名 I D と分岐順と空要素順の情報を少なくとも含み、  
前記属性出現情報は、着目属性の出現する文書番号と文字位置と祖先パス名 I D と要素名 I D と分岐順と空要素順の情報を少なくとも含み、

前記テキスト出現情報は、要素実体テキストおよび属性値から切り出された部分文字列に関し、出現する文書番号と文字位置と祖先パス名IDと要素名IDと属性名IDと分岐順と空要素順の情報を少なくとも含むことを特徴とする請求項3に記載のデータベース構築装置。

【請求項7】

前記祖先パス名登録部は、前記構造化文書に出現する各祖先パス名を1つ以上に分割した各々の部分祖先パス名に対してユニークな祖先パス名IDを割り当てて祖先パス名辞書に登録することを特徴とする請求項1に記載のデータベース構築装置。

【請求項8】

前記要素出現情報格納部に同じ要素名IDをキーにして登録されている前記要素出現情報のエントリ群と、前記祖先パス出現情報格納部に同じ祖先パス名IDをキーにして登録されている前記祖先パス出現情報のエントリ群とに対して、文書番号と文字位置以外の1つ以上の情報項目の値が共通するエントリ同士をグループ化する出現情報グループ化部を有することを特徴とする請求項1に記載のデータベース構築装置。

【請求項9】

構造化文書を管理するデータベース検索装置において、  
構造化文書に出現する各要素名に対してユニークな要素名IDを登録した要素名辞書と、  
前記構造化文書に出現する各祖先パス名に対してユニークな祖先パス名IDを登録した祖先パス名辞書と、  
前記構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名IDと分岐順の情報を少なくとも含む要素出現情報を、要素名IDをキーとして格納した要素出現情報格納部と、  
前記構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と要素名IDと分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名IDをキーとして格納した、祖先パス出現情報格納部と、  
検索式を入力するための検索条件入力部と、  
前記要素名辞書と前記祖先パス名辞書とを参照して、前記入力された検索式を内部条件式に変換する検索条件解析部と、  
前記検索条件解析部の出力した内部条件式にしたがって、前記要素出現情報格納部からの要素出現情報および、前記祖先パス出現情報格納部からの祖先パス出現情報から検索結果群を求める出現情報取得部と、  
を有することを特徴とするデータベース検索装置。

【請求項10】

属性名IDと対応する属性名の記録された属性名辞書と、  
着目属性の出現する文書番号と文字位置と祖先パス名IDと要素名IDと分岐順の情報を少なくとも含む属性出現情報を、属性名IDをキーとして格納した属性出現情報格納部とを有し、  
前記検索条件解析部が、前記要素名辞書と前記祖先パス名辞書と前記属性名辞書とを参照して、前記検索条件入力部から入力された検索式を内部条件式に変換し、前記出現情報取得部が、前記検索条件解析部の出力した内部条件式にしたがって、前記要素出現情報格納部からの要素出現情報、前記祖先パス出現情報格納部からの祖先パス出現情報および、前記属性出現情報格納部からの属性出現情報から検索結果群を求めることを特徴とする請求項9に記載のデータベース検索装置。

【請求項11】

要素実体テキストおよび属性値から切り出された部分文字列に関し、出現する文書番号と文字位置と祖先パス名IDと要素名IDと属性名IDと分岐順の情報を少なくとも含むテキスト出現情報を、切り出された部分文字列をキーとして格納した、テキスト出現情報格納部とを有し、  
前記出現情報取得部が、前記検索条件解析部の出力した内部条件式にしたがって、前記要素出現情報格納部からの要素出現情報、前記祖先パス出現情報格納部からの祖先パス出現

情報、前記属性出現情報格納部からの属性出現情報および、前記テキスト出現情報格納部からのテキスト出現情報から検索結果群を求めることを特徴とする請求項 9 に記載のデータベース検索装置。

【請求項 1 2】

前記出現情報取得部は、前記要素出現情報格納部における指定要素名 I D のエントリ数と、前記祖先パス出現情報格納部における指定祖先パス名 I D のエントリ数の大小を比較し、少ない方の出現情報を参照するようにして検索結果群を求めることを特徴とする請求項 9 乃至 1 1 のいずれかに記載のデータベース検索装置。

【請求項 1 3】

構造化文書を管理するデータベース構築方法において、  
構造化文書にユニークな文書番号を割り当てるとともに構造の解析を行うステップと、  
前記解析結果に基づいて、構造化文書に出現する各要素名に対してユニークな要素名 I D を割り当てて要素名辞書に登録するステップと、  
前記解析結果に基づいて、構造化文書に出現する各祖先パス名に対してユニークな祖先パス名 I D を割り当てて祖先パス名辞書に登録するステップと、  
前記解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名 I D と分岐順の情報を少なくとも含む要素出現情報を、要素名 I D をキーとして要素出現情報格納部に、着目要素の出現する文書番号と文字位置と要素名 I D と分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名 I D をキーとして祖先パス出現情報格納部にそれぞれ登録するステップと、を有することを特徴とするデータベース構築方法。

【請求項 1 4】

前記要素出現情報は、着目要素の出現する文書番号と文字位置と祖先パス名 I D と分岐順と空要素順の情報を少なくとも含み、  
前記祖先パス出現情報は、着目要素の出現する文書番号と文字位置と要素名 I D と分岐順と空要素順の情報を少なくとも含むことを特徴とする、請求項 1 3 に記載のデータベース構築方法。

【請求項 1 5】

前記祖先パス名辞書に登録するステップは、構造化文書に出現する各祖先パス名を 1 つ以上に分割した各々の部分祖先パス名に対してユニークな祖先パス名 I D を割り当てて登録するステップであり、  
前記要素出現情報には、単一の祖先パス名 I D の代わりに 1 つ以上の祖先パス名 I D の列を含み、  
前記祖先パス出現情報格納部には、単一の祖先パス名 I D の代わりに 1 つ以上の祖先パス名 I D の列をキーとして前記祖先パス出現情報を登録することを特徴とする、請求項 1 3 に記載のデータベース構築方法。

【請求項 1 6】

前記要素出現情報格納部に同一の要素名 I D をキーとして登録され、文書番号と文字位置以外の情報項目の値が共通であるような前記要素出現情報のエントリ同士をグループ化し、前記祖先パス出現情報格納部に同一の祖先パス名 I D をキーとして登録され、文書番号と文字位置以外の情報項目の値が共通であるような前記祖先パス出現情報のエントリ同士をグループ化するステップを有することを特徴とする請求項 1 3 に記載のデータベース構築方法。

【請求項 1 7】

構造化文書を管理するデータベース検索方法において、  
構造化文書に出現する各要素名に対してユニークな要素名 I D を登録した要素名辞書と、  
前記構造化文書に出現する各祖先パス名に対してユニークな祖先パス名 I D を登録した祖先パス名辞書と、  
前記構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名 I D と分岐順の情報を少なくとも含む要素出現情報を、要素名 I D をキーとして格納した要素出現情報格納部と、

前記構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と要素名 I D と分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名 I D をキーとして格納した、祖先パス出現情報格納部と、  
検索式を入力するためのステップと、  
前記要素名辞書と前記祖先パス名辞書とを参照して、前記入力された検索式を内部条件式に変換するステップと、  
前記内部条件式にしたがって、前記要素出現情報格納部からの要素出現情報および、前記祖先パス出現情報格納部からの祖先パス出現情報から検索結果群を求めるステップと、  
を有することを特徴とするデータベース検索方法。

【請求項 18】

構造化文書を管理するデータベース装置において、  
構造化文書に出現する各要素名に対してユニークな要素名 I D を記憶する要素名辞書と、  
前記構造化文書に出現する各祖先パス名に対してユニークな祖先パス名 I D を記憶する祖先パス名辞書と、  
構造化文書にユニークな文書番号を割り当てるとともに構造の解析を行う入力文書解析部と、  
前記入力文書解析部の解析結果に基づいて、前記構造化文書に出現する各要素名に対してユニークな要素名 I D を割り当てて前記要素名辞書に登録する要素名登録部と、  
前記入力文書解析部の解析結果に基づいて、前記構造化文書に出現する各祖先パス名に対してユニークな祖先パス名 I D を割り当てて前記祖先パス名辞書に登録する祖先パス名登録部と、  
文書番号と文字位置と祖先パス名 I D と分岐順の情報を少なくとも含む要素出現情報を、要素名 I D をキーとして記憶する要素出現情報格納部と、  
文書番号と文字位置と要素名 I D と分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名 I D をキーとして記憶する祖先パス出現情報格納部と、  
前記入力文書解析部の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名 I D と分岐順の情報を少なくとも含む要素出現情報を、前記着目要素の要素名 I D をキーとして前記要素出現情報格納部に登録し、かつ、前記着目要素の出現する文書番号と文字位置と要素名 I D と分岐順の情報を少なくとも含む祖先パス出現情報を、前記着目要素の祖先パス名 I D をキーとして前記祖先パス出現情報格納部に登録する出現情報登録部とを具備するデータベース構築装置と、  
検索式を入力する検索条件入力部と、  
前記要素名辞書と前記祖先パス名辞書とを参照して、前記検索条件入力部で入力された検索式について要素名と祖先パス名とをそれぞれ要素名 I D と祖先パス名 I D とで表現した内部条件式に変換する検索条件解析部と、  
前記要素出現情報格納部に記憶している要素出現情報、および、前記祖先パス出現情報格納部に記憶している祖先パス出現情報から、前記検索条件解析部で生成された前記内部条件式にあてはまる検索結果群データを抽出する出現情報取得部とを具備するデータベース検索装置と  
を有することを特徴とするデータベース装置。

【請求項 19】

属性名 I D と対応する属性名を記憶する属性名辞書と、  
前記入力文書解析部の解析結果に基づいて、前記構造化文書に出現する各属性名に対してユニークな属性名 I D を割り当てて前記属性名辞書に登録する属性名登録部と、  
文書番号と文字位置と祖先パス名 I D と要素名 I D と分岐順の情報を少なくとも含む属性出現情報を、属性名 I D をキーとして記憶する属性出現情報格納部とをさらに有し、  
前記出現情報登録部は、さらに、前記入力文書解析部の解析結果に基づいて、着目属性の出現する文書番号と文字位置と祖先パス名 I D と要素名 I D と分岐順の情報を少なくとも含む属性出現情報を、属性名 I D をキーとして前記属性出現情報格納部に登録するようにし、

前記検索条件解析部は、さらに、前記属性名辞書を参照して、前記検索条件入力部で入力された検索式について、属性名を属性IDで表現した内部条件式に変換するようにし、前記出現情報取得部は、さらに、前記要素出現情報格納部に記憶している要素出現情報と、前記祖先パス出現情報格納部に記憶している祖先パス出現情報と、前記属性出現情報格納部に記憶している属性出現情報とから前記検索条件解析部の出力した前記内部条件式にあてはまる検索結果群データを抽出することを特徴とする請求項18に記載のデータベース装置。



【書類名】 明細書

【発明の名称】 データベース構築装置及びデータベース検索装置及びデータベース装置

【技術分野】

【0001】

本発明は、XMLなどの論理構造を有する構造化文書を管理するデータベース装置に関し、特に、大量の構造化文書を蓄積管理するデータベース構築装置とそれに蓄積された構造化文書を効率良く検索するデータベース検索装置に関する。

【背景技術】

【0002】

構造化文書を論理構造に基づいて登録し、論理構造を指定した全文検索をする装置として、構造化文書管理装置が知られている（例えば、特許文献1参照）。

【0003】

以下、従来例の概要について図を参照しながら説明する。図33は、従来の構造化文書管理装置の構成図である。登録対象の構造化文書は構造化文書入力部2402から入力し、構造解析部2407によって解析され、木構造を得る。構造情報作成部2408によって、各要素のタグ名（要素名）には名称IDが割り振られて名称IDテーブル格納部2418に格納される。また、各要素のパス名称（最上位階層から順にタグ名を連ねて記述した文字列）には、パス名称IDが割り振られて、パス名称インデックス格納部2416に格納されている。各要素のパス階層（パス名称の各階層の出現順序（同じ親要素を持つ同じタグ名の要素の中で何番目に出現した要素か）を連ねて記述した文字列）には、パス階層IDが割り当てられて、パス階層インデックス格納部2417に格納されている。実体（テキスト）を持つ要素（要素実体）の場合は、各要素実体に対し、検索単位を一意に表す符合（検索単位識別子と呼ぶ）が割り当てられ、この検索単位識別子をキーとして、文書番号、パス名称ID、パス階層ID、名称IDの組が要素管理テーブル格納部2415に格納される。図34は、従来の構造化文書管理装置における要素管理テーブルの例を示す図である。図34においては、要素管理テーブル格納部2415に格納される要素管理テーブルの例を示したものである。

【0004】

次に、文字列索引作成部2409は、各要素実体の内容の文字列に対して、予め定めた文字数の文字連鎖を取り出す。この文字連鎖について、該当する検索単位識別子、および該文字連鎖先頭文字がその要素内容において何番目の文字かを表す番号（文字位置番号）を文字列索引格納部2419に登録する。図35は、従来の構造化文書管理装置における文字列索引の例の一部を示す図である。図35において、2601は「検索単位識別子が“1”の要素の文字列中に“構造”という文字連鎖が先頭から“1”文字目の位置から存在する」ということを表している。

【0005】

次に、このようにして格納されたデータを用いた検索の概要を説明する。図36は検索条件として「パス名称が“／論文／書誌／タイトル”である要素に“構造化”という文字列が含まれる文書」が与えられた場合の処理を図に示したものである。検索条件解析部2410は、パス名称インデックス2416を参照し、検索条件のパス名称をパス名称ID“N2”に変換する。次に文字列索引検索部2411は“構造化”から2文字連鎖“構造”と“造化”を取り出す。文字列索引を参照し、“構造”と“造化”が連続して出現し、かつ検索単位識別子が同一なものを求め、その検索単位識別子を抽出する。図36は、従来の構造化文書管理装置における検索処理を説明する図である。図36において、検索単位識別子“1”と“8”が文字列索引検索結果群として返っている。次に、構造照合部2412が検索条件の構造指定を満たす最終的な検索結果を求める。文字列索引検索結果群として得られた検索単位識別子をキーにして、要素管理テーブルを参照し、パス名称IDが“N2”に一致するものだけを最終的な検索結果とする。

【0006】

その他、タグ名を指定した検索条件であれば、要素管理テーブルの名称IDが指定タグ

名の名称 I D と一致するものだけを最終的な検索結果とする。また、パス名称とパス階層をともに指定した検索条件であれば、要素管理テーブルのパス名称 I D が指定したパス名称のパス名称 I D と一致し、かつパス階層 I D が指定したパス階層のパス階層 I D と一致するものだけを最終的な検索結果とする。

#### 【 0 0 0 7 】

また、別の文書管理装置として、構造化文書に含まれる要素を階層構造上の位置と結び付けるインデックスを生成し、階層構造上の位置までの探索経路が同じである要素（すなわち 1 の親ノードに対して複数の子ノードが存在するような構成）であっても複数の要素それぞれを識別するよう管理する文書管理装置が知られている（例えば、特許文献 2 参照）。

【特許文献 1】特開 2 0 0 2 - 2 0 2 9 7 3 号公報（第 2 2 頁、第 1 図）

【特許文献 2】特開 2 0 0 4 - 3 1 0 6 0 7 号公報（第 1 4 頁、第 1 図）

#### 【発明の開示】

#### 【発明が解決しようとする課題】

#### 【 0 0 0 8 】

しかしながら、上記従来の構造化文書管理装置では、まず文字列索引を参照して指定された文字列の出現する検索単位識別子を求めた後、検索単位識別子が指定された構造条件を満たすかどうかを、要素管理テーブルを参照して判定するため、文字列検索条件の指定は必須であり、構造条件だけを指定した検索を行うことができない。すなわち、検索を行うためには全ての検索単位識別子について構造条件を満たすかどうかを判定しなければならないため、効率が非常に悪いという課題がある。また、構造化文書データを蓄積する際に、全文検索のための検索インデックスデータに論理構造データを付加する構造としているため、そのような構造条件だけを指定した検索に対して効率的な検索を可能とする構造の検索用データを構築することができないという課題がある。

#### 【 0 0 0 9 】

また、文字列索引は要素実体の内容文字列に対してのみ作成されるため、要素の属性値に対しては文字列検索を行うことができないという課題がある。

#### 【 0 0 1 0 】

本発明は、このような課題を解決するもので、文字列検索条件と構造条件をともに指定した場合だけでなく、文字列検索条件を伴わない構造だけを指定した様々な検索条件に対しても、所望の文書を効率良く検索することが可能な構造の検索用データを構築し、効率良く検索可能なデータベース装置を提供することを目的とする。

#### 【 0 0 1 1 】

また、本発明は、要素内のテキスト文字列だけでなく、属性値に対しても文字列検索が可能な検索用データを構築し、効率良く検索可能なデータベース装置を提供することを目的とする。

#### 【課題を解決するための手段】

#### 【 0 0 1 2 】

前記従来の課題を解決するために、本発明のデータベース構築装置は、構造化文書にユニークな文書番号を割り当てるとともに構造の解析を行う入力文書解析部と、入力文書解析部の解析結果に基づいて、構造化文書に出現する各要素名に対してユニークな要素名 I D を割り当てて要素名辞書に登録する要素名登録部と、入力文書解析部の解析結果に基づいて、構造化文書に出現する各祖先パス名に対してユニークな祖先パス名 I D を割り当てて祖先パス名辞書に登録する祖先パス名登録部と、入力文書解析部の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名 I D と分岐順の情報を少なくとも含む要素出現情報を、要素名 I D をキーとして要素出現情報格納部に登録し、かつ、文書番号と文字位置と要素名 I D と分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名 I D をキーとして祖先パス出現情報格納部に登録する出現情報登録部とを備える。

#### 【 0 0 1 3 】

そのため、構造化文書を登録蓄積する際に、要素の出現情報に基づいて適切な出現情報インデックスを生成し、文字列検索条件と構造条件をともに指定した場合だけでなく、文字列検索条件を伴わない構造条件だけを指定した様々な検索条件に対しても、所望の文書を効率良く検索することが可能な構造の検索用データを構築することができる。

#### 【 0 0 1 4 】

また、本発明のデータベース構築装置は、入力文書解析部の解析結果に基づいて、構造化文書に出現する各属性名に対してユニークな属性名 I D を割り当てて属性名辞書に登録する属性名登録部を有し、出現情報登録部が、入力文書解析部の解析結果に基づいて、着目属性の出現する文書番号と文字位置と祖先パス名 I D と要素名 I D と分岐順の情報を少なくとも含む属性出現情報を、属性名 I D をキーとして属性出現情報格納部に登録する。

#### 【 0 0 1 5 】

そのため、構造化文書の登録の際に、属性に関する構造情報を登録できるようになり、結果として属性に関する構造条件を指定して所望の文書を効率良く検索することが可能な構造の検索用データを構築することができる。

#### 【 0 0 1 6 】

また、本発明のデータベース構築装置は、出現情報登録部が、入力文書解析部の解析結果に基づいて、要素実体テキストおよび属性値から切り出された部分文字列に関し、出現する文書番号と文字位置と祖先パス名 I D と要素名 I D と属性名 I D と分岐順の情報を少なくとも含むテキスト出現情報を、切り出された部分文字列をキーとしてテキスト出現情報格納部に登録する。

#### 【 0 0 1 7 】

そのため、構造化文書の登録の際に、要素実体テキストおよび属性値の部分文字列に関する構造情報を登録できるようになり、結果として要素実体テキストおよび属性値の部分文字列に関する構造条件を指定して所望の文書を効率良く検索することが可能な構造の検索用データを構築することができる。

#### 【 0 0 1 8 】

また、本発明のデータベース構築装置は、要素出現情報は、着目要素の出現する文書番号と文字位置と祖先パス名 I D と分岐順と空要素順の情報を少なくとも含み、祖先パス出現情報は、着目要素の出現する文書番号と文字位置と要素名 I D と分岐順と空要素順の情報を少なくとも含む。

#### 【 0 0 1 9 】

そのため、構造化文書の登録の際に、要素が要素実体のテキストを全く含まない要素（空要素）に関する構造情報を登録できるようになり、結果として空要素に関する構造条件を指定して所望の文書を効率良く検索することが可能な構造の検索用データを構築することができる。

#### 【 0 0 2 0 】

また、本発明のデータベース構築装置は、要素出現情報は、着目要素の出現する文書番号と文字位置と祖先パス名 I D と分岐順と空要素順の情報を少なくとも含み、祖先パス出現情報は、着目要素の出現する文書番号と文字位置と要素名 I D と分岐順と空要素順の情報を少なくとも含み、属性出現情報は、着目属性の出現する文書番号と文字位置と祖先パス名 I D と要素名 I D と分岐順と空要素順の情報を少なくとも含む。

#### 【 0 0 2 1 】

そのため、構造化文書の登録の際に、属性がテキストを全く含まない要素（空要素）に関する構造情報を登録できるようになり、結果として属性の空要素に関する構造条件を指定して所望の文書を効率良く検索することが可能な構造の検索用データを構築することができる。

#### 【 0 0 2 2 】

また、本発明のデータベース構築装置は、要素出現情報は、着目要素の出現する文書番号と文字位置と祖先パス名 I D と分岐順と空要素順の情報を少なくとも含み、祖先パス出現情報は、着目要素の出現する文書番号と文字位置と要素名 I D と分岐順と空要素順の情

報を少なくとも含み、属性出現情報は、着目属性の出現する文書番号と文字位置と祖先パス名IDと要素名IDと分岐順と空要素順の情報を少なくとも含み、テキスト出現情報は、要素実体テキストおよび属性値から切り出された部分文字列に関し、出現する文書番号と文字位置と祖先パス名IDと要素名IDと属性名IDと分岐順と空要素順の情報を少なくとも含む。

#### 【0023】

そのため、構造化文書の登録の際に、要素実体テキストおよび属性値から切り出された部分文字列がテキストを全く含まない要素（空要素）に関する構造情報を登録できるようになり、結果として要素実体テキストおよび属性値から切り出された部分文字列の空要素に関する構造条件を指定して所望の文書を効率良く検索することが可能な構造の検索用データを構築することができる。

#### 【0024】

また、本発明のデータベース構築装置は、祖先パス名登録部は、構造化文書に出現する各祖先パス名を1つ以上に分割した各々の部分祖先パス名に対してユニークな祖先パス名IDを割り当てて祖先パス名辞書に登録する。

#### 【0025】

そのため、構造化文書の登録の際に、祖先パス名を分割して部分パスを重複して蓄積しないように祖先パス列として登録できるようになり、結果として祖先パス辞書のサイズが小さく、構造条件を指定して所望の文書を効率良く検索することが可能な構造の検索用データを構築することができる。

#### 【0026】

また、本発明のデータベース構築装置は、要素出現情報格納部に同じ要素名IDをキーにして登録されている要素出現情報のエントリ群と、祖先パス出現情報格納部に同じ祖先パス名IDをキーにして登録されている祖先パス出現情報のエントリ群とに対して、文書番号と文字位置以外の1つ以上の情報項目の値が共通するエントリ同士をグループ化する出現情報グループ化部を備える。

#### 【0027】

そのため、登録されている構造化文書の出現位置情報の共通する値の項目を重複して蓄積しないようにグループ化して登録できるようになり、結果として出現位置索引のサイズが小さく、構造条件を指定して所望の文書を効率良く検索することが可能な構造の検索用データを構築することができる。

#### 【0028】

また、本発明のデータベース検索装置は、構造化文書に出現する各要素名に対してユニークな要素名IDを登録した要素名辞書と、構造化文書に出現する各祖先パス名に対してユニークな祖先パス名IDを登録した祖先パス名辞書と、構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名IDと分岐順の情報を少なくとも含む要素出現情報を、要素名IDをキーとして格納した要素出現情報格納部と、構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と要素名IDと分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名IDをキーとして格納した、祖先パス出現情報格納部と、検索式を入力するための検索条件入力部と、要素名辞書と祖先パス名辞書とを参照して、入力された検索式を内部条件式に変換する検索条件解析部と、検索条件解析部の出力した内部条件式にしたがって、要素出現情報格納部からの要素出現情報および、祖先パス出現情報格納部からの祖先パス出現情報から検索結果群を求める出現情報取得部とを備える。

#### 【0029】

そのため、構造化文書を検索する際に、要素と祖先パスの出現情報に基づく適切な出現情報インデックスを参照できるようになり、結果として文字列検索条件を伴わない要素名と祖先パス名に関する構造条件だけを指定した検索条件に対して所望の構造化文書を効率良く検索することができる。

#### 【0030】

また、本発明のデータベース検索装置は、属性名IDと対応する属性名の記録された属性名辞書と、着目属性の出現する文書番号と文字位置と祖先パス名IDと要素名IDと分岐順の情報を少なくとも含む属性出現情報を、属性名IDをキーとして格納した属性出現情報格納部とを有し、検索条件解析部が、要素名辞書と祖先パス名辞書と属性名辞書とを参照して、検索条件入力部から入力された検索式を内部条件式に変換し、出現情報取得部が、検索条件解析部の出力した内部条件式にしたがって、要素出現情報格納部からの要素出現情報、祖先パス出現情報格納部からの祖先パス出現情報および、属性出現情報格納部からの属性出現情報から検索結果群を求める。

#### 【0031】

そのため、構造化文書を検索する際に、要素名と祖先パス名と属性名に関する出現情報インデックスを参照できるようになり、結果としてそれらに関する構造条件だけを指定した検索条件に対して所望の構造化文書を効率良く検索することができる。

#### 【0032】

また、本発明のデータベース検索装置は、要素実体テキストおよび属性値から切り出された部分文字列に関し、出現する文書番号と文字位置と祖先パス名IDと要素名IDと属性名IDと分岐順の情報を少なくとも含むテキスト出現情報を、切り出された部分文字列をキーとして格納した、テキスト出現情報格納部とを有し、出現情報取得部が、検索条件解析部の出力した内部条件式にしたがって、要素出現情報格納部からの要素出現情報、祖先パス出現情報格納部からの祖先パス出現情報、属性出現情報格納部からの属性出現情報および、テキスト出現情報格納部からのテキスト出現情報から検索結果群を求める。

#### 【0033】

そのため、構造化文書を検索する際に、要素名と祖先パス名と属性名と要素実体テキストおよび属性値から切り出された部分文字列に関する出現情報インデックスを参照できるようになり、結果としてそれらに関する構造条件だけを指定した検索条件に対して所望の構造化文書を効率良く検索することができる。

#### 【0034】

また、本発明のデータベース検索装置は、出現情報取得部は、要素出現情報格納部における指定要素名IDのエントリ数と、祖先パス出現情報格納部における指定祖先パス名IDのエントリ数の大小を比較し、少ない方の出現情報を参照するようにして検索結果群を求める。

#### 【0035】

そのため、構造化文書を検索する際に、構造化文書に含まれる論理構造の要素数に応じて少ないエントリの出現情報を選択できるようになり、結果として検索対象が出現するエントリ数の絞込みが速く、構造条件だけを指定した検索条件に対して所望の構造化文書を効率良く検索することができる。

#### 【0036】

また、本発明のデータベース装置は、構造化文書に出現する各要素名に対してユニークな要素名IDを記憶する要素名辞書と、構造化文書に出現する各祖先パス名に対してユニークな祖先パス名IDを記憶する祖先パス名辞書と、構造化文書にユニークな文書番号を割り当てるとともに構造の解析を行う入力文書解析部と、入力文書解析部の解析結果に基づいて、構造化文書に出現する各要素名に対してユニークな要素名IDを割り当てて要素名辞書に登録する要素名登録部と、入力文書解析部の解析結果に基づいて、構造化文書に出現する各祖先パス名に対してユニークな祖先パス名IDを割り当てて祖先パス名辞書に登録する祖先パス名登録部と、文書番号と文字位置と祖先パス名IDと分岐順の情報を少なくとも含む要素出現情報を、要素名IDをキーとして記憶する要素出現情報格納部と、文書番号と文字位置と要素名IDと分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名IDをキーとして記憶する祖先パス出現情報格納部と、入力文書解析部の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名IDと分岐順の情報を少なくとも含む要素出現情報を、着目要素の要素名IDをキーとして要素出現情報格納部に登録し、かつ、着目要素の出現する文書番号と文字位置と要素名IDと分岐順の情

報を少なくとも含む祖先パス出現情報を、着目要素の祖先パス名IDをキーとして祖先パス出現情報格納部に登録する出現情報登録部とを具備するデータベース構築装置と、検索式を入力する検索条件入力部と、要素名辞書と祖先パス名辞書とを参照して、検索条件入力部で入力された検索式について要素名と祖先パス名とをそれぞれ要素名IDと祖先パス名IDとで表現した内部条件式に変換する検索条件解析部と、要素出現情報格納部に記憶している要素出現情報、および、祖先パス出現情報格納部に記憶している祖先パス出現情報から、検索条件解析部で生成された内部条件式にあてはまる検索結果群データを抽出する出現情報取得部とを具備するデータベース検索装置とを備える。

#### 【0037】

そのため、要素の出現情報に基づいて適切な出現情報インデックスを生成し、文字列検索条件と構造条件をともに指定した場合だけでなく、文字列検索条件を伴わない構造条件だけを指定した様々な検索条件に対しても、所望の文書を効率良く検索することが可能な構造の検索用データを構築し、また、効率良く検索することができる。

#### 【0038】

また、本発明のデータベース装置は、属性名IDと対応する属性名を記憶する属性名辞書と、入力文書解析部の解析結果に基づいて、構造化文書に出現する各属性名に対してユニークな属性名IDを割り当てて属性名辞書に登録する属性名登録部と、文書番号と文字位置と祖先パス名IDと要素名IDと分岐順の情報を少なくとも含む属性出現情報を、属性名IDをキーとして記憶する属性出現情報格納部とをさらに有し、出現情報登録部は、さらに、入力文書解析部の解析結果に基づいて、着目属性の出現する文書番号と文字位置と祖先パス名IDと要素名IDと分岐順の情報を少なくとも含む属性出現情報を、属性名IDをキーとして属性出現情報格納部に登録するようにし、検索条件解析部は、さらに、属性名辞書を参照して、検索条件入力部で入力された検索式について、属性名を属性IDで表現した内部条件式に変換するようにし、出現情報取得部は、さらに、要素出現情報格納部に記憶している要素出現情報と、祖先パス出現情報格納部に記憶している祖先パス出現情報と、属性出現情報格納部に記憶している属性出現情報とから検索条件解析部の出力した内部条件式にあてはまる検索結果群データを抽出する。

#### 【0039】

そのため、構造化文書の登録の際に、属性に関する構造情報を登録できるようになり、結果として属性に関する構造条件を指定して所望の文書を効率良く検索することが可能な構造の検索用データを構築することができ、また、効率良く検索することができる。

#### 【発明の効果】

#### 【0040】

本発明のデータベース装置によれば、文字列検索条件と構造条件をともに指定した検索条件のみならず、構造だけを指定した様々な検索条件に対しても、所望の論理構造を持つ文書を効率良く検索するデータベースが構築でき、さらに効率良く検索することが可能となる。

#### 【0041】

また、要素実体のテキスト文字列に対してだけでなく、属性値に対しても文字列検索を行うことが可能となる。

#### 【発明を実施するための最良の形態】

#### 【0042】

以下、本発明の実施の形態におけるデータベース装置について、図面を参照しながら説明する。

#### 【0043】

##### （実施の形態1）

本実施の形態におけるデータベース装置の構成および動作について説明する。図1は、本発明の実施の形態1におけるデータベース装置の構成を示すブロック図である。図1において、101はデータベースに登録する構造化文書群、102は入力された構造化文書群101の各文書についてユニークな文書番号を割り振るとともに論理構造の解析を行う

入力文書解析部、103は入力文書解析部102の解析結果から、文書に出現する要素名に対してユニークな識別子（以下、要素名IDと呼ぶ）を割り当てて要素名辞書107に登録する要素名登録部、104は入力文書解析部102の解析結果から、文書に出現する祖先パス名（着目要素の祖先要素の要素名を最上位階層から順にスラッシュで区切って並べた文字列で、着目要素自身の要素名は含まない）に対してユニークな識別子（以下、祖先パス名IDと呼ぶ）を割り当てて祖先パス名辞書108に登録する祖先パス名登録部、105は入力文書解析部102の解析結果から、文書に出現する属性名に対してユニークな識別子（以下、属性名IDと呼ぶ）を割り当てて属性名辞書109に登録する属性名登録部、106は入力文書解析部102の解析結果から、出現位置索引110の要素出現情報格納部111、祖先パス出現情報格納部112、属性出現情報格納部113、テキスト出現情報格納部114に4種の出現情報を登録する出現情報登録部、107は要素名IDとそれに対応する要素名が記録された要素名辞書、108は祖先パス名IDとそれに対応する祖先パス名が記録された祖先パス名辞書、109は属性名IDとそれに対応する属性名が記録された属性名辞書、110は要素出現情報格納部111、祖先パス出現情報格納部112、属性出現情報格納部113、テキスト出現情報格納部114、の4種の出現情報が格納されている出現位置索引格納部、111は各要素の出現する文書番号、文字位置、文字数、祖先パス名ID、分岐順の情報を、要素名IDをキーにして格納した要素出現情報格納部、112は各要素の出現する文書番号、文字位置、文字数、要素名ID、分岐順の情報を、その要素の祖先パス名IDをキーにして格納した、祖先パス出現情報格納部、113は各属性の出現する文書番号、文字位置、文字数、要素名ID、祖先パス名ID、分岐順の情報を、属性名IDをキーにして格納した属性出現情報格納部、114は要素内のテキストから切り出した部分文字列、および要素の持つ属性の値から切り出した部分文字列に関して、出現する文書番号、文字位置、祖先パス名ID、要素名ID、属性名ID、分岐順の情報を、部分文字列をキーにして格納したテキスト出現情報格納部、116は検索式115を受け付ける検索条件入力部、117は、検索条件入力部116に与えられた検索式を解析し、内部条件に変換して出現情報取得部118に出力する検索条件解析部、118は検索条件解析部117の出力した内部条件にしたがって、出現位置索引110に格納された4種の出現情報から適切な情報を選択して取得し、検索条件にマッチする結果データ集合を求める出現情報取得部、119は結果データ集合を適切な形式で検索結果120として出力する検索結果出力部である。

#### 【0044】

上記のように構成されたデータベース装置の動作について説明する。はじめに、文書登録（データベース構築）処理に関して具体例を挙げて説明する。図2は、本発明の実施の形態1における文書登録処理の手順を示す流れ図である。

#### 【0045】

まず、ステップ2201において、入力文書解析部102は、構造化文書群101から構造化文書を1つ読み込んで、ユニークな文書番号を割り振る。

#### 【0046】

次に、ステップ2202において、入力文書解析部102は、この文書の論理構造を解析する。図3は、本発明の実施の形態1における登録検索対象となる構造化文書の一例を示す図である。構造化文書群101には、このような図3に示す文書が複数含まれる。図3に示した構造化文書は、最上位階層にbook要素を持ち、book要素はtitle要素と2つのchapter要素を含んでいる。title要素は、要素実体の文字列“文書検索”を含み、1つ目のchapter要素は別のtitle要素と2つのsection要素および属性値が“歴史”であるkeyword属性を持つ構造を持っている。図3に示す構造化文書を入力文書解析部102によって解析した結果得られる木構造は、図4のようになる。図4は、本発明の実施の形態1における構造化文書の論理構造を解析した結果である木構造の一例を示す図である。図4において、四角い枠は要素301～303を表し、枠内に記された文字列は要素名304を示している。また、楕円の点線枠は属性305を表し、枠内に記された文字列は属性名306を示している。木構造の最上位

階層の要素 3 0 1 から着目要素に至る経路の途中に存在する要素（祖先要素）の要素名をスラッシュで区切って順に並べたものはパス名と呼ばれる。パス名のうちの末尾部分（＝着目要素自身の要素名）を除いた部分を「祖先パス名」と呼ぶことにする。図 5 は、本発明の実施の形態 1 における祖先パス名を説明する図である。図 5 において、図 4 の網掛けを施した要素 3 0 2 に関するパス名 7 0 1、祖先パス名 7 0 2、要素名 7 0 3 を示している。

#### 【 0 0 4 7 】

また、図 4 において、要素の右肩に記された “ 1 / 2 / 3 ” などの文字列は、パス名中の各要素について、同じ親要素を持つ同じ要素名の要素の中で何番目に出現したかの順を示す番号を並べたもので、これを「分岐順」3 0 7 と呼ぶ。図 4 の網掛けを施した要素 3 0 2 とその左隣の要素 3 0 3 とは、パス名は同じであるが分岐順 3 0 7、3 0 8 は異なっている。なお、分岐順の表記方法はこれに限らない。例えば、1 以外の値を持つ階層の深さとその値を並べる方法でもよい。分岐順 3 0 7 ( “ 1 / 2 / 3 ” ) をこの方法で表記すれば、深さ 1 の値は 1 なので省略、深さ 2 の値が 2、深さ 3 の値が 3、したがって “ 2 : 2, 3 : 3 ” となる。同じ要素名の兄弟要素がめったに現れない文書、すなわち、分岐順の値がほとんど 1 であるような文書を格納する場合には、このような表記方法の方が出現位置索引ファイルのサイズを小さくできる。

#### 【 0 0 4 8 】

次に、入力文書解析部 1 0 2 の解析結果をうけて、当該文書に出現する各要素について以下の処理を繰り返す。

#### 【 0 0 4 9 】

ステップ 2 2 0 3 において、要素名登録部 1 0 3 は、着目要素の要素名が要素名辞書 1 0 7 に登録済みかどうかを調べ、登録済みであれば対応する要素名 I D を取得し、登録されていない場合は新たに要素名 I D ( > 0 ) を割り当てて要素名辞書 1 0 7 に登録する。

#### 【 0 0 5 0 】

ステップ 2 2 0 4 において、祖先パス名登録部 1 0 4 は、着目要素の祖先パス名が祖先パス名辞書 1 0 8 に登録済みかどうかを調べ、登録済みであれば対応する祖先パス名 I D を取得し、登録されていない場合は新たに祖先パス名 I D ( > 0 ) を割り当てて祖先パス名辞書 1 0 8 に登録する。

#### 【 0 0 5 1 】

もし、着目要素が属性を持っているならば、ステップ 2 2 0 5 ～ステップ 2 2 0 6 において、属性名登録部 1 0 5 は、着目要素の各属性の属性名が属性名辞書 1 0 9 に登録済みかどうかを調べ、登録済みであれば対応する属性名 I D を取得し、登録されていない場合は新たに属性名 I D ( > 0 ) を割り当てて属性名辞書 1 0 9 に登録する。図 6 は、本発明の実施の形態 1 における要素名辞書の内容の一例を示す図である。また、図 7 は、本発明の実施の形態 1 における祖先パス名辞書の内容の一例を示す図である。また、図 8 は、本発明の実施の形態 1 における属性名辞書の内容の一例を示す図である。図 7、図 8、図 9 において、それぞれ構造化文書（図 3）の登録処理が終わった後の要素名辞書 1 0 7、祖先パス名辞書 1 0 8、属性名辞書 1 0 9 の内容の例を示している。

#### 【 0 0 5 2 】

ステップ 2 2 0 7 において、出現情報登録部 1 0 6 は、着目要素に関する要素出現情報を、要素名 I D をキーとして要素出現情報格納部 1 1 1 に登録する。要素出現情報は、文書番号、着目要素（子孫要素も含む）に含まれる（タグ以外の）テキストの先頭文字位置および文字数、祖先パス名 I D、分岐順の 5 種類の値の組から構成される。なお、「文字位置」は、図 9 に示すように、タグを除く当該文書内の全てのテキストをつなげた文字列において先頭から何文字目にあたるかで表す。また、着目要素が要素実体のテキストを全く含まない要素（＝空要素）である場合には、着目要素以降に初めて現れる（タグ以外の）テキストの先頭文字位置を着目要素の先頭文字位置とみなす。図 1 0 は、本発明の実施の形態 1 における要素出現情報を説明する図である。図 1 0 において、図 4 の網掛けを施した要素 3 0 2 に関する要素出現情報が、要素名 I D が 4 （＝要素名が s e c t i o n）



である要素が文書番号1の文書の115文字目から始まる長さ40文字の要素実体を含んでいて、その祖先パス名IDが3（＝祖先パス名が／book／chapter）で分岐順が1／2／3であることを表している。

#### 【0053】

ステップ2208において、出現情報登録部106は、着目要素に関する祖先パス出現情報（すなわち、文書番号、着目要素（子孫要素も含む）に含まれる（タグ以外の）テキストの先頭文字位置および文字数、要素名ID、分岐順の5種類の値の組）を、祖先パス名IDをキーとして祖先パス出現情報格納部112に登録する。図11は、本発明の実施の形態1における祖先パス出現情報を説明する図である。図11において、図4の網掛けを施した要素302に関する祖先パス出現情報の内容を示している。図10と図11を比較してわかるように、同一要素に関する要素出現情報と祖先パス出現情報は、キーとなる項目が要素名IDであるか祖先パス名IDであるかという点が異なるだけである。

#### 【0054】

もし、着目要素が属性を持っているならば、ステップ2209～ステップ2210において、出現情報登録部106は着目要素の各属性に関する属性出現情報を、属性名IDをキーとして属性出現情報格納部113に登録する。属性出現情報は、文書番号、属性値の先頭文字位置および文字数、祖先パス名ID、要素名ID、分岐順の6種類の値の組から構成される。図12は、本発明の実施の形態1における属性出現情報を説明する図である。図12において、図4の網掛けを施した要素302の「update」属性305に関する属性出現情報の内容を示している。その内容は、属性名IDが2（＝属性名がupdate）の属性が文書番号1の文書の115文字目から始まる長さ6文字の属性値を持ち、属性の所属する要素の祖先パス名IDが3（＝祖先パス名が／book／section）、要素名IDが4（＝要素名がsection）、分岐順が1／2／3であることを示している。なお、属性出現情報において、属性値の先頭文字位置は、図12に示すように、仮想的に着目要素（子孫要素も含む）に含まれる（タグ以外の）テキストの先頭文字位置と同じであるとする。

#### 【0055】

ステップ2211において、出現情報登録部106は、着目要素の実体内容のテキストから部分文字列の切り出しを行い、テキスト出現情報を、切り出された部分文字列をキーとしてテキスト出現情報格納部114に登録する。ただし、属性値ではないので、属性名IDには常に0を格納する。テキスト出現情報は、文書番号、切り出された部分文字列の先頭文字位置、祖先パス名ID、要素名ID、属性名ID、分岐順の6種類の値の組から構成される。

#### 【0056】

もし、着目要素が属性を持っているならば、ステップ2212～ステップ2213において、出現情報登録部106は、着目要素が持つ各属性の属性値文字列から部分文字列の切り出しを行い、テキスト出現情報格納部114に部分文字列をキーとして登録する。なお、属性出現情報と同様に、属性値は図11に示すような位置に仮想的に出現しているとして、文字位置を算出する。また、ステップ2213ではステップ2211の場合とは異なり、属性名IDには着目している属性の属性名ID（＞0）を格納する。図13は、本発明の実施の形態1におけるテキスト出現情報を説明する図である。図13において、図4の網掛けを施した要素302のテキストおよび「update」属性305の属性値についてのテキスト出現情報の一部である。図13において、1201は、“極大”という部分文字列が文書番号1の文書の118文字目に現れ、祖先パス名IDが3（＝祖先パス名が／book／section）、要素名IDが4（要素名がchapter）、分岐順が1／2／3であるような要素の要素実体に含まれている（属性名IDが0であることからわかる）ことを表している。また1202は、“00”という部分文字列が文書番号1の文書の116文字目に現れ、祖先パス名IDが3（＝祖先パス名が／book／section）、要素名IDが4（＝要素名がchapter）、分岐順が1／2／3であるような要素に属する属性名IDが2（＝属性名がupdate）の属性の属性値に含ま

れていることを表している。

#### 【0057】

ステップ2214において、この文書に出現する全ての要素について処理が終わったかどうかを調べ、もし未処理の要素が残っていればステップ2203に戻って処理を繰り返す。

#### 【0058】

ステップ2215において、全ての入力文書に対して処理が終わったかどうかを調べ、未処理の文書が残っていればステップ2201に戻って処理を繰り返す。

#### 【0059】

以上のようにして、文書登録（データベース構築）処理が完了する。

#### 【0060】

続いて、登録済みの文書群に対する検索処理に関して説明する。図14は、本発明の実施の形態1における検索式の例を示す図である。図14においては、検索条件入力部116に与えられる検索式115の例をいくつか示したもので、これらの式はW3C（World Wide Web Consortium）の勧告として公開されているXPath言語（詳細な仕様は<http://www.w3.org/TR/xpath>に記載されている）で記述されている。

#### 【0061】

図14のそれぞれのXPath式は、次のような意味を表している。検索式2101は「最上位階層のbook要素の子のchapter要素の子であるtitle要素」を表している。検索式2102は「最上位階層のbook要素の子のchapter要素のいずれかの子要素」を表している。検索式2103は、「いずれかの階層にあるtitle要素」を表している。検索式2104は「最上位階層のbook要素の子のchapter要素の子の2番目のsection要素」を表している。検索式2105は、「最上位階層のbook要素の子のchapter要素の子のsection要素のupdate属性」を表している。検索式2106は、「最上位階層のbook要素の子のchapter要素の子のsection要素で、かつ要素実体内容に“極大単語”という文字列を含む要素」を表している。検索式2107は、「最上位階層のbook要素の子のchapter要素の子のsection要素のupdate属性で、かつその属性値に“2004”という文字列を含む」を表している。

#### 【0062】

次に、それぞれの検索式に対して、本実施の形態におけるデータベース装置でどのような検索処理が行われるのかを順に説明する。図15は、本発明の実施の形態1におけるデータベース装置の検索処理の手順を示す流れ図である。

#### 【0063】

（検索式2101の場合）

図15に沿って、検索式2101の場合の検索処理の流れを説明する。

ステップ2301において、検索条件入力部116に入力された検索式2101は、検索条件解析部117で解析される。

#### 【0064】

ステップ2302において、検索条件解析部117は、検索式2101を解析し、要素名辞書107、祖先パス名辞書108を参照して内部条件「祖先パス名ID=3かつ要素名ID=2」に変換し、出現情報取得部118に出力する。

#### 【0065】

次に、ステップ2303からステップ2305において、出現情報取得部118は、出現位置索引110を参照し、要素出現情報格納部111における要素名ID=2のエントリ数Nと祖先パス出現情報格納部112における祖先パス名ID=3のエントリ数Mとを比較し、少ない方を選択する。図16は、要素出現情報格納部111における要素名ID=2のエントリ1301、図17は祖先パス出現情報格納部112における祖先パス名ID=3のエントリ1401の例で、この場合はN=8、M=12であるから図16の要素

出現情報格納部 1 1 1 を選ぶことになる。

【 0 0 6 6 】

そして、ステップ 2 3 0 6 において、出現情報取得部 1 1 8 は、要素出現情報格納部 1 1 1 の要素名 I D = 2 のエントリ 1 3 0 1 から 1 つ取得し、ステップ 2 3 0 7 で、このエントリの祖先パス名 I D が 3 であるかどうかを調べ、もし祖先パス名 I D が 3 であればステップ 2 3 0 8 でこのエントリのデータを結果データ集合 1 3 0 2 に追加する。結果データ集合の各データは例えば（文書番号，祖先パス名 I D，要素名 I D，属性名 I D，分岐順）のような形式である。

【 0 0 6 7 】

ステップ 2 3 0 9 において、出現情報取得部 1 1 8 は、N エントリ全てについて処理したか調べ、まだ未処理のエントリがあればステップ 2 3 0 6 に戻って処理を繰り返す。

【 0 0 6 8 】

ステップ 2 3 0 5 において、出現情報取得部 1 1 8 は、もし  $M \leq N$  であれば、図 1 7 のように祖先パス出現情報格納部 1 1 2 における祖先パス名 I D = 3 の各エントリ 1 4 0 1 を調べ、要素名 I D が 2 であるものを求め（ステップ 2 3 1 0 ～ステップ 2 3 1 3）結果データ集合 1 4 0 2 に追加する。

【 0 0 6 9 】

ステップ 2 3 1 4 において、出現情報取得部 1 1 8 は、求められた結果データ集合を検索結果出力部 1 1 9 に出力する。

【 0 0 7 0 】

最後に検索結果出力部 1 1 9 は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

【 0 0 7 1 】

このように、検索式 2 1 0 1 に対しては、要素出現情報格納部 1 1 1 における指定した要素名 I D のエントリから指定した祖先パス名 I D を持つものを選ぶという処理と、祖先パス出現情報格納部 1 1 2 における指定した祖先パス名 I D のエントリから指定した要素名 I D を持つものを選ぶという 2 種類の処理のどちらか、エントリ数の少ない方を選ぶことによって、検索対象構造化文書群の論理構造の特性に応じて処理量を抑えることができるため、所望の文書を効率良く検索することができる。

【 0 0 7 2 】

（検索式 2 1 0 2 の場合）

検索条件入力部 1 1 6 に入力された検索式 2 1 0 2 は、検索条件解析部 1 1 7 で解析される。検索条件解析部 1 1 7 は、検索式 2 1 0 2 を解析し、祖先パス名辞書 1 0 8 を参照して内部条件「祖先パス名 I D = 3」に変換し、出現情報取得部 1 1 8 に出力する。出現情報取得部 1 1 8 は、出現位置索引 1 1 0 を参照し、図 1 8 のように祖先パス出現情報格納部 1 1 2 における祖先パス名 I D = 3 の全てのエントリ 1 5 0 1 を求め、例えば（文書番号，祖先パス名 I D，要素名 I D，属性名 I D，分岐順）のような形式で結果データ集合 1 5 0 2 として検索結果出力部 1 1 9 に出力する。検索結果出力部 1 1 9 は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

【 0 0 7 3 】

このように、検索式 2 1 0 2 に対しては、祖先パス出現情報格納部 1 1 2 における指定した祖先パス名 I D のエントリを取得するだけで良いため、所望の文書を効率良く検索することができる。

【 0 0 7 4 】

（検索式 2 1 0 3 の場合）

検索条件入力部 1 1 6 に入力された検索式 2 1 0 3 は、検索条件解析部 1 1 7 で解析される。検索条件解析部 1 1 7 は、検索式 2 1 0 3 を解析し、要素名辞書 1 0 7 を参照して内部条件「要素名 I D = 2」に変換し、出現情報取得部 1 1 8 に出力する。出現情報取得部 1 1 8 は、出現位置索引 1 1 0 を参照し、図 1 9 のように要素出現情報格納部 1 1 1 における要素名 I D = 2 の全てのエントリ 1 6 0 1 を求め、例えば（文書番号，祖先パス名

I D，要素名 I D，属性名 I D，分岐順) のような形式で結果データ集合 1 6 0 2 を検索結果出力部 1 1 9 に出力する。検索結果出力部 1 1 9 は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

#### 【0075】

このように、検索式 2 1 0 3 に対しては、要素出現情報格納部 1 1 1 における指定した要素名 I D のエントリを取得するだけで良いため、所望の文書を効率良く検索することができる。

#### 【0076】

(検索式 2 1 0 4 の場合)

検索条件入力部 1 1 6 に入力された検索式 2 1 0 4 は、検索条件解析部 1 1 7 で解析される。検索条件解析部 1 1 7 は、検索式 2 1 0 4 を解析し、要素名辞書 1 0 7、祖先パス名辞書 1 0 8 を参照して内部条件「祖先パス名 I D = 3 かつ要素名 I D = 4 かつ分岐順 = \* / \* / 2」に変換し、出現情報取得部 1 1 8 に出力する。分岐順のアスタリスク「\*」の部分はどんな数字でもマッチすることを表す。出現情報取得部 1 1 8 は、出現位置索引 1 1 0 を参照し、要素出現情報格納部 1 1 1 における要素名 I D = 4 のエントリ数 N と祖先パス出現情報格納部 1 1 2 における祖先パス名 I D = 3 のエントリ数 M とを比較し、少ない方を選択する。

#### 【0077】

もし、 $M \leq N$  であれば、図 2 0 に示すように祖先パス出現情報格納部 1 1 2 における祖先パス名 I D = 3 の各エントリ 1 7 0 1 を調べ、要素名 I D が 4 であり、かつ分岐順が「\* / \* / 2」であるエントリのデータを結果データ集合 1 7 0 2 として、例えば(文書番号，祖先パス名 I D，要素名 I D，属性名 I D，分岐順) のような形式で検索結果出力部 1 1 9 に出力する。もし、 $M > N$  ならば要素出現情報格納部 1 1 1 における要素名 I D = 4 の各エントリを調べ、祖先パス名 I D が 3 であり、かつ分岐順が「\* / \* / 2」であるエントリのデータを結果データ集合 1 7 0 2 として検索結果出力部 1 1 9 に出力する。

#### 【0078】

最後に検索結果出力部 1 1 9 は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

#### 【0079】

このように、検索式 2 1 0 4 に対しては、要素出現情報格納部 1 1 1 における指定した要素名 I D のエントリから指定した祖先パス名 I D と分岐順を持つものを選ぶという処理と、祖先パス出現情報格納部 1 1 2 における指定した祖先パス名 I D のエントリから指定した要素名 I D と分岐順を持つものを選ぶという2種類の処理のどちらか、エントリ数の少ない方を選ぶ。このことによって、処理量を減らすことが可能となり、所望の文書を効率良く検索することができる。

#### 【0080】

(検索式 2 1 0 5 の場合)

検索条件入力部 1 1 6 に入力された検索式 2 1 0 5 は、検索条件解析部 1 1 7 で解析される。検索条件解析部 1 1 7 は、検索式 2 1 0 5 を解析し、要素名辞書 1 0 7、祖先パス名辞書 1 0 8、属性名辞書 1 0 9 を参照して内部条件「祖先パス名 I D = 3 かつ要素名 I D = 4 かつ属性名 I D = 2」に変換し、出現情報取得部 1 1 8 に出力する。出現情報取得部 1 1 8 は、出現位置索引 1 1 0 を参照し、図 2 1 のように属性出現情報格納部 1 1 3 における属性名 I D = 2 の各エントリ 1 8 0 1 を調べ、祖先パス名 I D が 3 であり、要素名 I D が 4 であればそのエントリのデータを例えば(文書番号，祖先パス名 I D，要素名 I D，属性名 I D，分岐順) のような形式で結果データ集合 1 8 0 2 として検索結果出力部 1 1 9 に出力する。最後に、検索結果出力部 1 1 9 は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

#### 【0081】

このように、検索式 2 1 0 5 に対しては、属性出現情報格納部 1 1 3 における指定した属性名 I D のエントリから指定した祖先パス名 I D と要素名 I D を持つものを選ぶことに

よって、所望の文書を検索することが可能となる。

#### 【0082】

（検索式2106の場合）

検索条件入力部116に入力された検索式2106は、検索条件解析部117で解析される。検索条件解析部117は、検索式2106を解析し、要素名辞書107、祖先パス名辞書108を参照して内部条件「祖先パス名ID=3かつ要素名ID=4かつ要素内に文字列“極大単語”を含む」に変換し、出現情報取得部118に出力する。出現情報取得部118は、出現位置索引110を参照し、図22のようにテキスト出現情報格納部114における“極大”のエントリ1901と“単語”のエントリ1902の間の接続演算を行う。その際、文書番号が同一であることと“単語”が“極大”の2文字後方に位置することだけでなく、祖先パス名IDが3、かつ要素名IDが4、かつ属性名IDが0、かつ分岐順が同一であるというチェックも行い条件を満たすものを出力する。例えば（文書番号、祖先パス名ID、要素名ID、属性名ID、分岐順）のような形式で結果データ集合1903として検索結果出力部119に出力する。検索結果出力部119は、求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

#### 【0083】

このように、検索式2106に対しては、テキスト出現情報格納部114における部分文字列のエントリ同士の接続演算の際に、祖先パス名IDおよび要素名IDが指定した値であって、分岐順が同一であり、かつ属性名IDが0であるものを選ぶことによって、所望の文書を検索することが可能となる。

#### 【0084】

（検索式2107の場合）

検索条件入力部116に入力された検索式2107は、検索条件解析部117で解析される。検索条件解析部117は、検索式2107を解析し、要素名辞書107、祖先パス名辞書108、属性名辞書109を参照して内部条件「祖先パス名ID=3かつ要素名ID=4かつ属性名ID=2かつ属性値に文字列“2004”を含む」に変換し、出現情報取得部118に出力する。出現情報取得部118は、出現位置索引110を参照し、図23のようにテキスト出現情報格納部114における“20”のエントリ2001と“04”のエントリ2002の間の接続演算を行う。その際、文書番号が同一であることと“20”が“04”の2文字後方に位置することだけでなく、祖先パス名IDが3、かつ要素名IDが4、かつ属性名IDが2、かつ分岐順が同一であるというチェックも行い、条件を満たすものを出力する。例えば（文書番号、祖先パス名ID、要素名ID、属性名ID、分岐順）のような形式で結果データ集合2003として検索結果出力部119に出力する。検索結果出力部119は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

#### 【0085】

このように、検索式2107に対しては、テキスト出現情報格納部114における部分文字列のエントリ同士の接続演算の際に、祖先パス名IDおよび要素名IDが指定した値であって、分岐順が同一であり、かつ属性名IDが指定した値(>0)であるものを選ぶことによって、所望の文書を検索することが可能となる。

#### 【0086】

以上説明したように、要素の出現情報を、要素名IDをキーにして格納した要素出現情報格納部と、要素の出現情報をその要素の祖先パス名IDをキーにして格納した祖先パス出現情報格納部と、属性の出現情報を、属性名IDをキーにして格納した属性出現情報格納部とを設けることにより、構造条件だけを指定した検索式に対しても効率良く所望の文書を検索することができる。また、要素実体のテキスト文字列および要素の持つ属性の属性値から切り出された部分文字列の出現情報を格納したテキスト出現情報格納部を設けることにより、要素実体のテキストに対してだけでなく属性値に対しても文字列検索を行うことができる。

#### 【0087】

なお、データベース構築処理において、要素実体や属性値から固定長の2文字連鎖で部分文字列の切り出しを行うと説明したが、他の切り出し方法、例えば特開平8-249354号公報「文書検索装置および単語索引作成方法および文書検索方法」に記載の方法等でも構わない。

#### 【0088】

また、データベース検索処理において、検索条件式をXPath式で与えるとして説明したが、同様の意味を持つ他のクエリ言語であっても本発明を適用することは可能である。

#### 【0089】

このような構成とすることによって、本実施の形態では、構造化文書の登録の際に、構造化文書に含まれる文書構造を示す要素名と祖先パス名と属性名の一覧と、それらの構造化文書中での出現位置情報のインデックスを生成することにより、構造化文書構造の全文検索のみならず、文書構造を示す検索式に示される文書を効率的に検索することができる。

#### 【0090】

なお、本実施の形態では、構造化文書を登録する際に、文書構造を解析して辞書データおよび出現位置索引データを構築して構造化文書を登録する構成と、受け付けた文書構造を示す検索式に示される文書を辞書データおよび出現位置索引データに基づいて登録文書を効率的に検索する構成とを同時に実現する形態としたが、登録する機能のみの構成、あるいは検索のみする構成として実現してもよい。

#### 【0091】

なお、本実施の形態では、構造化文書を登録する際に、要素と祖先パスに対する辞書データならびに出現位置索引データを生成して登録する構成と、この構成に属性に対する辞書データならびに出現位置索引データを生成して登録する構成と、さらにこの構成に要素や属性値のテキストに対する出現位置索引データを生成して登録する構成とを同時に実現する形態としたが、要素と祖先パスのみを対象として登録する構成、あるいは、この構成に属性を対象に加えて登録する構成、あるいは、さらにこの構成にテキストを対象に加えて登録する構成として実現してもよい。

#### 【0092】

##### （実施の形態2）

次に、本実施の形態2におけるデータベース装置の構成および動作について説明する。本実施の形態におけるデータベース装置の構成は、図1に示した実施の形態1と同じである。ただし、祖先パス登録部104が、文書に出現する各祖先パス名に対してではなく、祖先パス名をいくつかに分割した各部分祖先パス名に対してユニークな祖先パス名IDを割り当てて祖先パス名辞書108に登録すること、出現情報登録部106が、各要素の出現する文書番号、文字位置、文字数、祖先パス名ID列、分岐順、空要素順の情報を、要素名IDをキーにして要素出現情報格納部111へ、各要素の出現する文書番号、文字位置、文字数、要素名ID、分岐順、空要素順の情報を、祖先パス名ID列をキーにして祖先パス出現情報格納部112へ、各属性の出現する文書番号、文字位置、文字数、要素名ID、祖先パス名ID列、分岐順、空要素順の情報を、属性名IDをキーにして属性出現情報格納部113へ、要素内のテキストから切り出した部分文字列、および要素の持つ属性の値から切り出した部分文字列に関して、出現する文書番号、文字位置、祖先パス名ID列、要素名ID、属性名ID、分岐順、空要素順の情報を、部分文字列をキーにしてテキスト出現情報格納部114へそれぞれ格納する、という点が実施の形態1とは異なっている。

#### 【0093】

はじめに、文書登録（データベース構築）処理の動作について図2を用いて説明する。なお、実施の形態1と同様の処理を行う部分については詳細な説明を省略する。

#### 【0094】

ステップ2201において、入力文書解析部102は構造化文書を1つ読み込みユニークな文書番号を割り振った後、ステップ2202で、この構造化文書の論理構造を解析す

る。その際、実施の形態1の場合の処理に加え、各要素に関する「空要素順」の情報についても求める。「空要素順」とは、同じ親要素を持つ兄弟要素のうちで、先頭の要素であるかもしくは直前の兄弟要素が空要素（子孫要素を含めて要素実体のテキストを全く持たない要素）でない要素の場合には1、それ以外の場合（すなわち、直前の兄弟要素が空要素である場合）には、直前の兄弟要素の空要素順の値に1を加えた値を、最上位階層から当該要素に至るまでの各階層において求め並べたものである。

#### 【0095】

図24は、本発明の実施の形態2における空要素順の説明する図である。図24において、文書の木構造と空要素順の一例を示している。また、斜線模様の四角い枠は要素実体のテキストを含む要素2801、2804、2805を、無地の四角い枠は要素実体を含まない空要素2802、2803を、各要素の右肩に記された“1/2/3”のような文字列は、各要素の空要素順2806の情報を表している。

#### 【0096】

兄弟要素2801～2804の空要素順の最初の2つの数字“1/2”は祖先要素の空要素順にあたる部分で兄弟要素に共通であり、末尾の数字nが各要素毎に変わりうる。要素2801は兄弟要素の中の先頭要素であるのでn=1、要素2802は直前の要素2801が空要素ではないのでn=1、要素2803は直前の要素2802が空要素なので1を加えてn=2、要素2804は直前の要素2803が空要素なのでさらに1を加えてn=3となる。したがって、兄弟要素2801～2804の空要素順はそれぞれ、“1/2/1”、“1/2/1”、“1/2/2”、“1/2/3”となる。なお、空要素順の表記方法はこれに限らない。例えば、1以外の値を持つ階層の深さとその値を並べる方法でもよく、そのような方法で空要素順2806（“1/2/3”）を表記すれば、深さ1の値は1なので省略、深さ2の値が2、深さ3の値が3、したがって“2:2, 3:3”となる。空要素がほとんど現れない文書、すなわち、空要素順の値がほとんど1である文書を扱う場合には、後者の表記方法の方が出現位置索引ファイルのサイズを小さくできる。

#### 【0097】

次に、入力文書解析部102の解析結果をうけて、当該文書に出現する各要素について以下の処理を繰り返す。

#### 【0098】

ステップ2203では実施の形態1と同様の処理を行う。

#### 【0099】

ステップ2204において、祖先パス名登録部104は、着目要素の祖先パス名を3階層毎に分割していき、分割後の各部分祖先パス名が祖先パス名辞書108に登録済みかどうかを調べ、登録済みであれば対応する祖先パス名IDを取得し、登録されていない場合は新たに祖先パス名ID(>0)を割り当てて祖先パス名辞書108に登録する。なお、祖先パス名の深さが3階層以下ならば、祖先パス名ID列は実施の形態1の場合と同じように単一の祖先パス名IDとなる。図25は、本発明の実施の形態2における祖先パス名と祖先パス名ID列の例を示す図である。図25において、祖先パス名2901と対応する祖先パス名ID列2902、および、祖先パス名辞書108の内容2903の例を示している。このように祖先パス名を分割して各部分祖先パス名に祖先パス名IDを割り当てることで、当該要素の祖先要素や他の要素の処理において登録済の祖先パス名IDを共用することができる。また、祖先パス名IDの異なり数を小さくでき、祖先パス名辞書108のサイズを小さくすることが可能となる。

#### 【0100】

なお、本実施例では祖先パス名を3階層毎に分割する例を示したが、分割の方法はこれに限らない。例えば4階層毎に分割したり、階層の深さによって分割幅を変化させたりするようにしても構わない。また、祖先パス名ID列の区切り文字として“:”を用いたが他の区切り文字でも構わない。

#### 【0101】

もし、着目要素が属性を持っているならば、ステップ2205～ステップ2206にお

いて実施の形態 1 と同様の処理を行う。

#### 【0102】

ステップ 2207 において、出現情報登録部 106 は、着目要素に関する要素出現情報を、要素名 ID をキーとして要素出現情報格納部 111 に登録する。要素出現情報は、文書番号、着目要素（子孫要素も含む）に含まれる（タグ以外の）テキストの先頭文字位置および文字数、祖先パス名 ID 列、分岐順、空要素順の 6 種類の値の組から構成される。なお、「文字位置」は、タグを除く当該文書内の全てのテキストをつなげた文字列において先頭から何文字目にあたるかで表す。また、着目要素が要素実体のテキストを全く含まない要素（＝空要素）である場合には、着目要素以降に初めて現れる（タグ以外の）テキストの先頭文字位置を着目要素の先頭文字位置とみなす。要素出現情報の一例を図 26 に示す。図 26 は、本発明の実施の形態 2 における要素出現情報を説明する図である。実施の形態 1 と異なるのは、要素出現情報に単一の祖先パス名 ID ではなく 1 つ以上の祖先パス名 ID を区切り文字で連ねた祖先パス名 ID 列が記録されることと、空要素順の情報が含まれることである。

#### 【0103】

ステップ 2208 において、出現情報登録部 106 は、着目要素に関する祖先パス出現情報（すなわち、文書番号、着目要素（子孫要素も含む）に含まれる（タグ以外の）テキストの先頭文字位置および文字数、要素名 ID、分岐順、空要素順の 6 種類の値の組）を、祖先パス名 ID 列をキーとして祖先パス出現情報格納部 112 に登録する。祖先パス出現情報の一例を図 27 に示す。図 27 は、本発明の実施の形態 2 における祖先パス出現情報を説明する図である。実施の形態 1 と異なるのは、祖先パス出現情報に空要素順の情報が含まれることと、単一の祖先パス名 ID ではなく 1 つ以上の祖先パス名 ID を区切り文字で連ねた祖先パス名 ID 列をキーとして祖先パス名出現情報が祖先パス出現情報格納部 112 登録されることである。

#### 【0104】

もし、着目要素が属性を持っているならば、ステップ 2209 ～ステップ 2210 において、出現情報登録部 106 は着目要素の各属性に関する属性出現情報を、属性名 ID をキーとして属性出現情報格納部 113 に登録する。属性出現情報は、文書番号、属性値の先頭文字位置および文字数、祖先パス名 ID 列、要素名 ID、分岐順、空要素順の 7 種類の値の組から構成される。実施の形態 1 と異なるのは、属性出現情報に単一の祖先パス名 ID ではなく 1 つ以上の祖先パス名 ID を区切り文字で連ねた祖先パス名 ID 列が記録されることと、空要素順の情報が含まれることである。

#### 【0105】

ステップ 2211 において、出現情報登録部 106 は、着目要素の実体内容のテキストから部分文字列の切り出しを行い、テキスト出現情報を、切り出された部分文字列をキーとしてテキスト出現情報格納部 114 に登録する。ただし、テキスト出現情報は属性値ではないので、属性名 ID には常に 0 を格納する。テキスト出現情報は、文書番号、切り出された部分文字列の先頭文字位置、祖先パス名 ID 列、要素名 ID、属性名 ID、分岐順、空要素順の 7 種類の値の組から構成される。実施の形態 1 と異なるのは、テキスト出現情報に単一の祖先パス名 ID ではなく 1 つ以上の祖先パス名 ID を区切り文字で連ねた祖先パス名 ID 列が記録されることと、空要素順の情報が含まれることである。

#### 【0106】

もし、着目要素が属性を持っているならば、ステップ 2212 ～ステップ 2213 において、出現情報登録部 106 は、着目要素が持つ各属性の属性値文字列から部分文字列の切り出しを行い、テキスト出現情報格納部 114 に部分文字列をキーとして登録する。ステップ 2211 と同様、実施の形態 1 と異なるのは、テキスト出現情報に単一の祖先パス名 ID ではなく 1 つ以上の祖先パス名 ID を区切り文字で連ねた祖先パス名 ID 列が記録されることと、空要素順の情報が含まれることである。

#### 【0107】

以降ステップ 2214 ～2215 の処理を実施の形態 1 と同様に行い、文書登録（デー



データベース構築) 処理が完了する。

#### 【0108】

続いて、登録済みの文書群に対する検索処理に関して説明する。実施の形態1で説明した検索式と同様の形式を持つ検索式での検索処理については、検索条件解析部117において、祖先パス名から祖先パス名IDを求めて内部条件に変換する処理を、祖先パス名から祖先パス名ID列を求めるように変更すればよい。すなわち、祖先パス名を3階層毎に分割し、祖先パス名辞書108を参照して分割後の各部分祖先パス名に対応する祖先パス名IDを求め、それらの祖先パス名IDを順に区切り文字で区切って並べ祖先パス名ID列を求める。祖先パス名ID列の形式は、文書登録処理の説明で図25に示した例と同様であり、祖先パス名の深さが3階層以下の場合には単一の祖先パス名IDとなる。また、これに伴い、実施の形態1では出現情報取得部118において祖先パス名IDで照合していた各種処理を、祖先パス名ID列で照合するように変更することで、検索結果を求めることができるようになる。

#### 【0109】

(検索式3201の場合)

図28は、本発明の実施の形態2における検索式の例を示す図である。図28に示すXPath式は「最上位階層のA要素の子のB要素の子のX要素の兄弟要素で、X要素より後ろに現れるY要素」を表している。検索条件入力部116に入力された検索式3201は、検索条件解析部117で解析される。検索条件解析部117は、検索式3201を解析し、要素名辞書107、祖先パス名辞書108を参照して内部条件に変換し、出現情報取得部118に出力する。ただし、内部条件は、「C1かつ(C2またはC3)、ただし、Cx: {祖先パス名ID=25かつ要素名ID=10}、Cy: {祖先パス名ID=25かつ要素名ID=14}、C1: {CxとCyの文書番号が同一で、かつ分岐順が末尾以外等しい}、C2: {CxよりCyの方が文字位置の値が大きい}、C3: {CxとCyの文字位置の値が等しく、かつCxよりCyの方が空要素順の末尾の値が大きい}」である。ここで、祖先パス名“/A/B”に対応する祖先パス名IDが25、要素名“X”に対応する要素名IDが10、要素名“Y”に対応する要素名IDが14である。条件C3が必要なのは、空要素とその直後に位置する要素では文字位置が同一になるため、前後関係を判断するために空要素順の値を比較しなければならないからである。

#### 【0110】

図29は、本発明の実施の形態2における検索動作を説明する図である。出現情報取得部118は、出現位置索引110を参照し、図29に示すように、祖先パス出現情報格納部112における祖先パス名ID=25のエントリで要素名ID=10であるもの(Cx)、および要素名ID=14であるもの(Cy)を求める。続いて、C1かつ(C2またはC3)を満たすようなCx、Cyのエントリの組3301、3302を求める。例えば、(文書番号、祖先パス名ID、要素名ID、属性名ID、分岐順、空要素順)のような形式で結果データ集合3303として検索結果出力部119に出力する。検索結果出力部119は、求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

#### 【0111】

なお、CxおよびCyのエントリを求める際に、祖先パス出現情報格納部112における指定祖先パス名IDのエントリ数と、要素出現情報格納部111における指定要素名IDのエントリ数を比較して少ない方を選択するようにすることも可能である。

#### 【0112】

このようにして、検索式3201に対しては、祖先パス出現情報格納部112または要素出現情報格納部111を参照して求めた2つの要素の出現位置が同じだった場合(すなわち2つの要素が、空要素とその直後の要素の関係にあった場合)に、空要素順の情報を比較することによって、前後関係の曖昧さを排除し正しい検索結果を求めることができるようになる。

#### 【0113】

以上説明したように、祖先パス名登録部104が祖先パス名を分割し、分割後の各部分祖先パス名に対してユニークな祖先パス名IDを割り当てて祖先パス名辞書108に登録することで、祖先パス名辞書のサイズを小さくすることが可能となる。また、出現情報登録部106が要素出現情報格納部111、祖先パス出現情報格納部112、属性出現情報格納部113、テキスト出現情報格納部114に空要素順の情報も格納することにより、空要素とその直後の要素の開始文字位置が同じになることによる前後関係の曖昧さを排除し、正しい検索結果を求めることができる。

#### 【0114】

このような構成とすることによって、本実施の形態では、構造化文書の要素にテキストが全く含まれない空要素である場合には、着目要素以降に初めて現れるテキストの先頭文字位置を着目要素の先頭文字位置とみなすものである。さらに空要素の出現順を出現位置インデクスとして生成することより、構造化文書に空要素が含まれる場合だけでなく空要素が連続して含まれる場合であっても、構造化文書構造の全文検索のみならず、空要素を含む文書構造を示す検索式に示される文書を効率的に検索することができる。また、本実施の形態におけるデータベース装置は、祖先パス名を一定の条件で分割した部分パス名に基づいて祖先パス列として登録することにより、部分パスを重複して蓄積することなく、結果的に祖先パス辞書のサイズを小さくでき、また、構造化対象を多く含む構造化文書であっても、文書構造を示す検索式に示される文書を効率的に検索することができる。

#### 【0115】

なお、本実施の形態では、構造化文書を登録する際に、文書構造を解析して辞書データおよび出現位置索引データを構築して構造化文書を登録する構成と、受け付けた文書構造を示す検索式に示される文書を辞書データおよび出現位置索引データに基づいて登録文書を効率的に検索する構成とを同時に実現する形態としたが、構造化文書を登録する機能のみの構成、あるいは検索のみする構成として実現してもよい。

#### 【0116】

なお、本実施の形態では、構造化文書を登録する際に、テキスト要素を持たない空要素に対応する出現位置索引データを生成して登録する構成と、祖先パス名をいくつかに分割した各部分祖先パス名に対する辞書データならびに出現位置索引データを生成して登録する構成とを同時に実現する形態としたが、空要素のみを対象として登録する構成、あるいは、祖先パス名のみを対象として登録する構成として実現してもよい。

#### 【0117】

（実施の形態3）

次に、本実施の形態3におけるデータベース装置の構成および動作について説明する。図30は、本発明の実施の形態3におけるデータベース装置の構成を示すブロック図である。図30において、要素出現情報格納部111、祖先パス出現情報格納部112、属性出現情報格納部113、テキスト出現情報格納部114に格納されている情報のグループ化を行う出現情報グループ化部3401が追加されている点が、実施の形態1および実施の形態2の構成とは異なる。

#### 【0118】

はじめに、文書登録（データベース構築）処理の動作について説明する。図31は、本発明の実施の形態3におけるデータベース装置の文書登録処理の手順を示す流れ図である。図31において、ステップ2201～2215までの処理は実施の形態2の場合と同じであるので、説明を省略する。

#### 【0119】

最後のステップ3501において、出現情報グループ化部3401は要素出現情報格納部111に同じ要素名IDをキーとして登録されているエントリ群の中で、文書番号と文字位置を除いた4種類の情報項目（文字数、祖先パス名ID、分岐順、空要素順）の値が全て共通しているようなエントリ同士を集め、それらのエントリの数が閾値（例えば10エントリ）を超えていたらそれらのエントリをグループ化する。次に、残ったエントリ群について、文書番号と文字位置を除いた4種類の情報項目（文字数、祖先パス名ID、分

岐順、空要素順）のうち、いずれか３種類の情報項目の値が共通しているエントリ群を求め、エントリの数が閾値を超えていたらグループ化する。なお、あるエントリが複数のグループに属する可能性があるが、その場合にはエントリ数の最も多いグループに入れるものとする。同様にしていずれか２種類の情報項目の値が共通するエントリのグループ、いずれか１種類の情報項目の値が共通するエントリのグループを順に作成し、残ったエントリは共通情報項目無しのグループとして登録する。

#### 【０１２０】

図３２は、本発明の実施の形態３におけるグループ化された要素出現情報を説明する図である。図３２において、グループ化された要素出現情報の例を示している。グループ情報３６０１～３６０４には、各グループに属するエントリに共通する情報項目の値が格納され、個々のエントリ３６０５～３６０８には、共通しない情報項目の値のみが格納されている。第１のグループ情報３６０１は、当該グループに属する要素出現情報のエントリはどれも（文字数＝１０，祖先パス名ＩＤ＝１００，分岐順＝“１／１／１”，空要素順＝“１／１／１”）という値を共通に持つということを表している。そして、当該グループに属する個々のエントリ３６０５にはそれぞれの文書番号と文字位置だけが格納されている。第２のグループ情報３６０２は、当該グループに属する要素出現情報のエントリはどれも（祖先パス名ＩＤ＝２００，分岐順＝“１／２／１”，空要素順＝“１／２／３”）という値を共通に持ち、“\*”となっている文字数の情報項目は共通な値ではないということを表している。そして、個々のエントリ３６０６に文書番号、文字位置とともに文字数が格納されている。同様に第３のグループ情報３６０３は、当該グループに属する要素出現情報のエントリはどれも（文字数＝８，祖先パス名ＩＤ＝１５０，空要素順＝“１／２”）という値を共通に持ち、“\*”となっている分岐順の情報項目は共通な値ではないということを表している。そして、個々のエントリ３６０７に文書番号、文字位置とともに分岐順が格納されている。最後のグループ情報３６０４は共通する情報項目がないグループで、各エントリ３６０８に全ての情報項目が格納されている。

#### 【０１２１】

祖先パス出現情報格納部１１２、属性出現情報格納部１１３、テキスト出現情報格納部１１４に格納されている各情報についても同様にして、文書番号と文字位置以外に共通な値の情報項目を持つエントリ同士のグループ化を行い、文書登録（データベース構築）処理が完了する。

#### 【０１２２】

登録済みの文書群に対する検索処理に関しては、グループ化された各エントリの内容とグループ情報から全ての情報項目の値を復元できるので、実施の形態１や実施の形態２と同様に検索結果を求めることができる。

#### 【０１２３】

このようにして、出現情報グループ化部３４０１を設け、出現位置索引１１０に格納されるエントリ群をグループ化し、グループ内で共通する情報項目の値を括りだし、個々のエントリには格納しないようにすることにより、索引サイズを減らすことが可能となる。

#### 【０１２４】

このような構成とすることによって、本実施の形態では、各要素、祖先パスなどの出現位置情報についてある条件下で情報項目の値が共通する部分をグループ化、共通化してない部分とは異なる構造で格納することによって、共通する部分を重複して蓄積することなく、結果的に索引のサイズを小さくできる。

#### 【産業上の利用可能性】

#### 【０１２５】

本発明に係るデータベース装置は、構造化文書を効率良く検索することが可能な構造の検索用データを構築し、効率良く検索可能なデータベース装置等に適している。

#### 【図面の簡単な説明】

#### 【０１２６】

【図１】 本発明の実施の形態１におけるデータベース装置の構成を示すブロック図

【図 2】 本発明の実施の形態 1 における文書登録処理の手順を示す流れ図

【図 3】 本発明の実施の形態 1 における登録検索対象となる構造化文書の一例を示す図

【図 4】 本発明の実施の形態 1 における構造化文書の論理構造を解析した結果である木構造の一例を示す図

【図 5】 本発明の実施の形態 1 における祖先パス名を説明する図

【図 6】 本発明の実施の形態 1 における要素名辞書の内容の一例を示す図

【図 7】 本発明の実施の形態 1 における祖先パス名辞書の内容の一例を示す図

【図 8】 本発明の実施の形態 1 における属性名辞書の内容の一例を示す図

【図 9】 本発明の実施の形態 1 における文字位置を説明する図

【図 10】 本発明の実施の形態 1 における要素出現情報を説明する図

【図 11】 本発明の実施の形態 1 における祖先パス出現情報を説明する図

【図 12】 本発明の実施の形態 1 における属性出現情報を説明する図

【図 13】 本発明の実施の形態 1 におけるテキスト出現情報を説明する図

【図 14】 本発明の実施の形態 1 における検索式の例を示す図

【図 15】 本発明の実施の形態 1 におけるデータベース装置の検索処理の手順を示す流れ図

【図 16】 本発明の実施の形態 1 におけるデータベース装置の検索動作を説明する図

【図 17】 本発明の実施の形態 1 におけるデータベース装置の検索動作を説明する図

【図 18】 本発明の実施の形態 1 におけるデータベース装置の検索動作を説明する図

【図 19】 本発明の実施の形態 1 におけるデータベース装置の検索動作を説明する図

【図 20】 本発明の実施の形態 1 におけるデータベース装置の検索動作を説明する図

【図 21】 本発明の実施の形態 1 におけるデータベース装置の検索動作を説明する図

【図 22】 本発明の実施の形態 1 におけるデータベース装置の検索動作を説明する図

【図 23】 本発明の実施の形態 1 におけるデータベース装置の検索動作を説明する図

【図 24】 本発明の実施の形態 2 における空要素順の説明に用いる図

【図 25】 本発明の実施の形態 2 における祖先パス名と祖先パス名 ID 列の例を示す図

【図 26】 本発明の実施の形態 2 における要素出現情報を説明する図

【図 27】 本発明の実施の形態 2 における祖先パス出現情報を説明する図

【図 28】 本発明の実施の形態 2 における検索式の例を示す図

【図 29】 本発明の実施の形態 2 における検索動作を説明する図

【図 30】 本発明の実施の形態 3 におけるデータベース装置の構成を示すブロック図

【図 31】 本発明の実施の形態 3 におけるデータベース装置の文書登録処理の手順を示す流れ図

【図 32】 本発明の実施の形態 3 におけるグループ化された要素出現情報を説明する図

【図 33】 従来の構造化文書管理装置の構成図

【図 34】 従来の構造化文書管理装置における要素管理テーブルの例を示す図

【図 35】 従来の構造化文書管理装置における文字列索引の例の一部を示す図

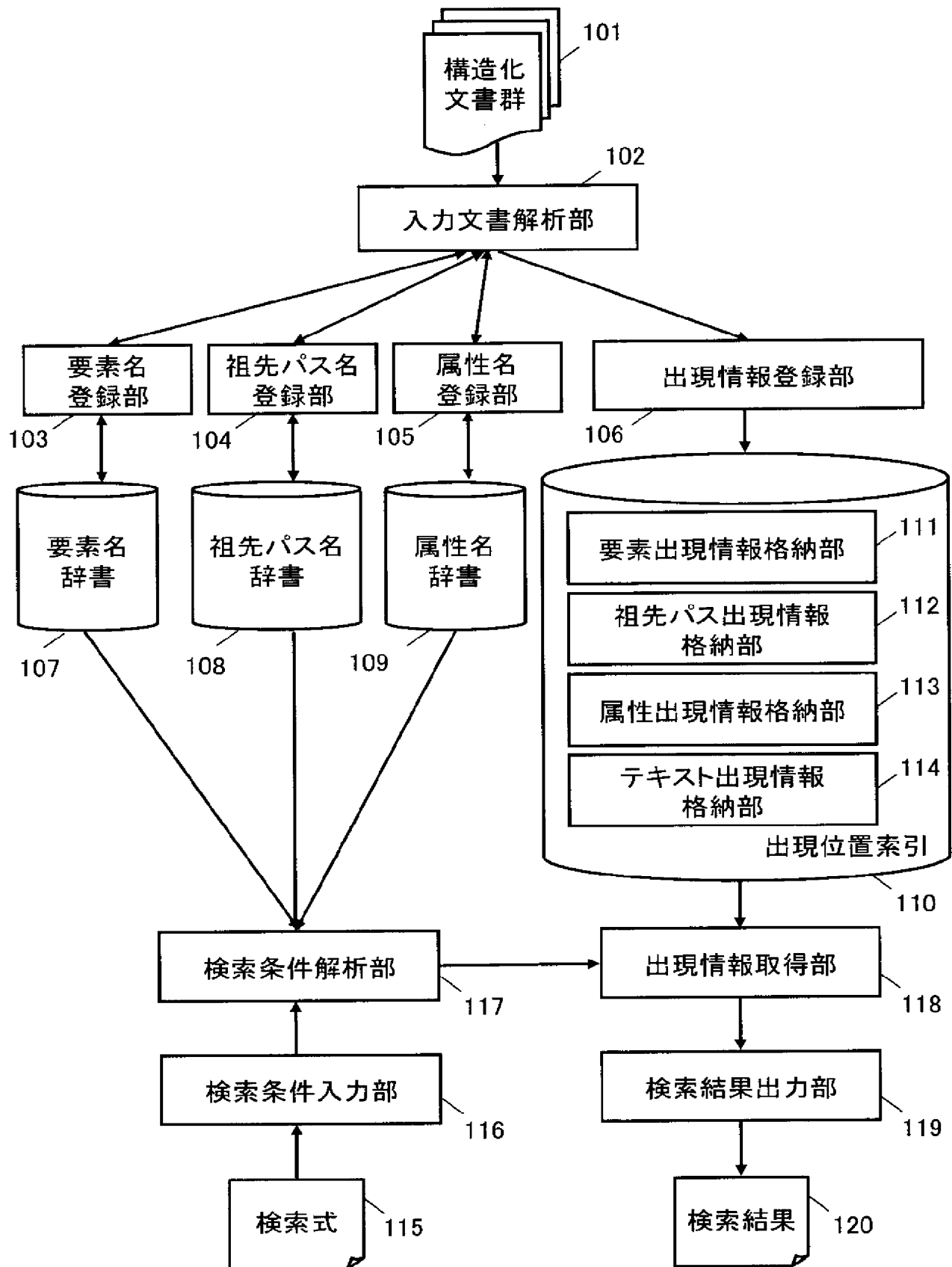
【図 36】 従来の構造化文書管理装置における検索処理を説明する図

## 【符号の説明】

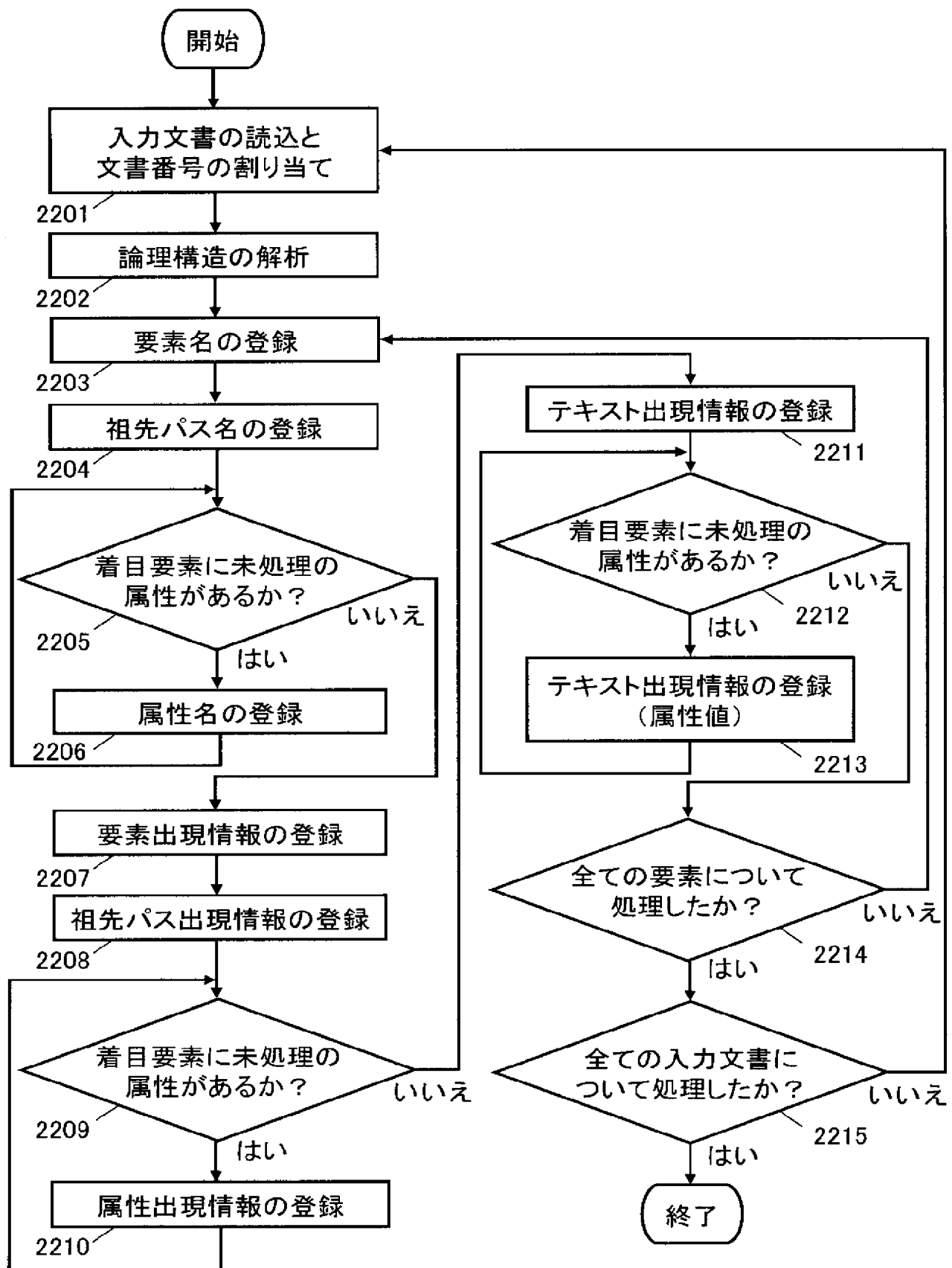
【0127】

- 101 構造化文書群
- 102 入力文書解析部
- 103 要素名登録部
- 104 祖先パス名登録部
- 105 属性名登録部
- 106 出現情報登録部
- 107 要素名辞書

1 0 8	祖先パス名辞書
1 0 9	属性名辞書
1 1 0	出現位置索引
1 1 1	要素出現情報格納部
1 1 2	祖先パス出現情報格納部
1 1 3	属性出現情報格納部
1 1 4	テキスト出現情報格納部
1 1 5	検索式
1 1 6	検索条件入力部
1 1 7	検索条件解析部
1 1 8	出現情報取得部
1 1 9	検索結果出力部
1 2 0	検索結果
3 4 0 1	出現情報グループ化部



【図 2】



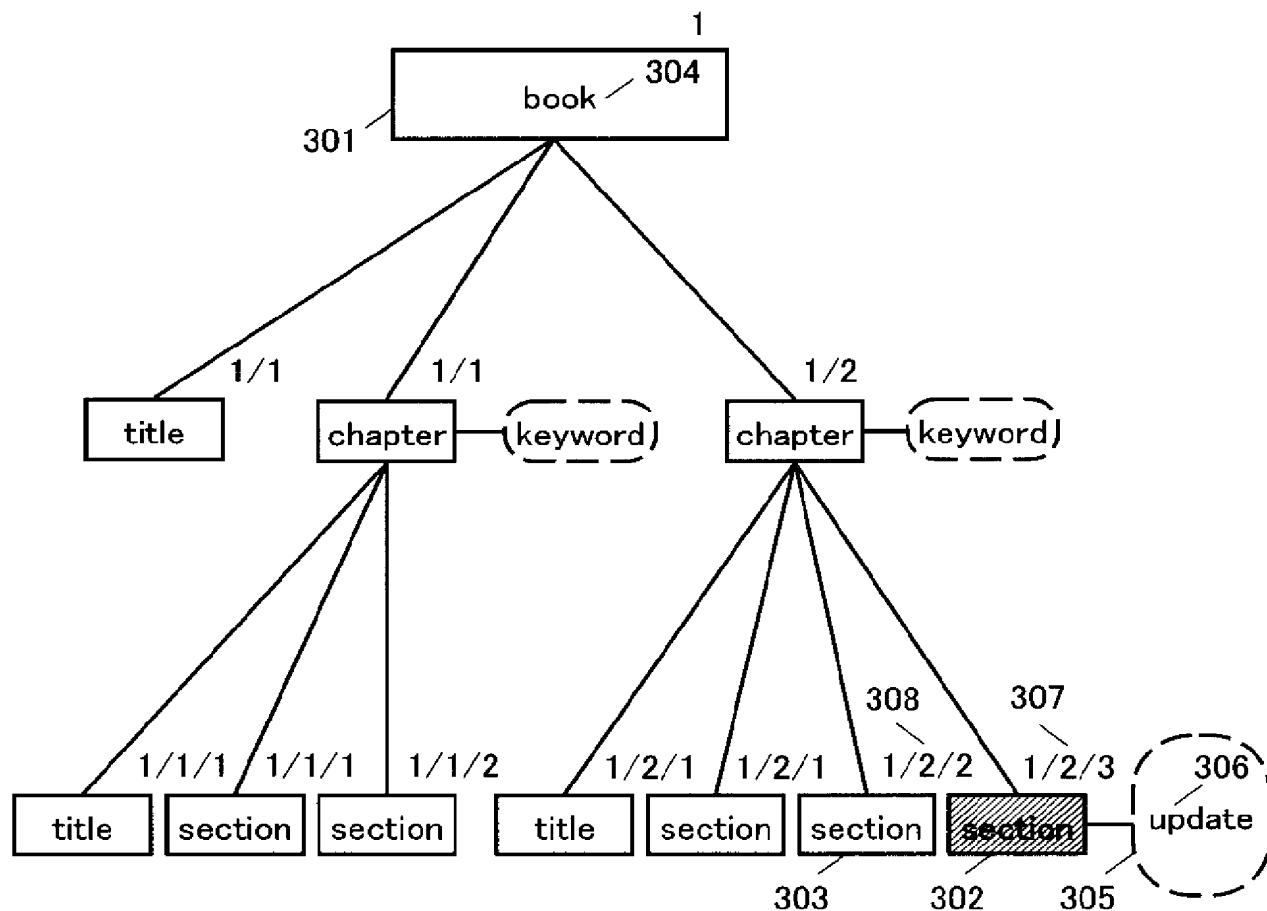
【図 3】

```

<book>
  <title>文書検索</title>
  <chapter keyword="歴史">
    <title>文書検索の歴史</title>
    <section>キーワード検索は、・・・</section>
    <section>その後、全文検索が・・・</section>
  </chapter>
  <chapter keyword="索引">
    <title>索引方式</title>
    <section>最長一致切り出しによる・・・</section>
    <section>n-gram索引方式は・・・</section>
    <section update="200406">新たに極大単語索引方式が・・・</section>
  </chapter>
</book>

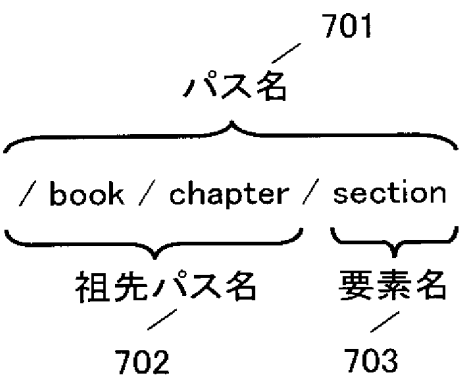
```

【図 4】





【図 5】



【図 6】

要素名ID	要素名
1	book
2	title
3	chapter
4	section

【図 7】

祖先パス名ID	祖先パス名
1	/
2	/book
3	/book/chapter

【図 8】

属性名ID	属性名
1	keyword
2	update

【図 9】

テキスト → 文 書 検 索 文 書 検 索 の 歴 史  
 文字位置 → 0 1 2 3 4 5 6 7 8 9 10

キ ー ワ ー ド 検 索 は 、 …  
 11 12 13 14 15 16 17 18 19

【図 10】

### 要素出現情報

要素名ID	文書番号	文字位置	文字数	祖先パス名 ID	分岐順
4	1	115	40	3	1/2/3

<section update="200406">新たに極大単語索引方式が…</section>

【図 1 1】

## 祖先パス出現情報

祖先パス名 ID	文書番号	文字位置	文字数	要素名ID	分岐順
3	1	115	40	4	1/2/3

【図 1 2】

## 属性出現情報

属性名ID	文書番号	文字位置	文字数	祖先パス名 ID	要素名ID	分岐順
2	1	115	6	3	4	1/2/3

200406

<section update="200406">新たに極大単語索引方式が・・・</section>

## テキスト出現情報

部分文字列	文書番号	文字位置	祖先パス名 ID	要素名ID	属性名ID	分岐順
“新た”	1	115	3	4	0	1/2/3
“たに”	1	116	3	4	0	1/2/3
“に極”	1	117	3	4	0	1/2/3
“極大”	1	118	3	4	0	1/2/3
“大単”	1	119	3	4	0	1/2/3
“単語”	1	120	3	4	0	1/2/3
<div> <div>1201</div> <div>属性値でない場合は0 属性値の場合は属性名ID(≠0)</div> </div>						
“20”	1	115	3	4	2	1/2/3
“00”	1	116	3	4	2	1/2/3
“04”	1	117	3	4	2	1/2/3
“40”	1	118	3	4	2	1/2/3
“06”	1	119	3	4	2	1/2/3

2101 / book / chapter / title

2102 / book / chapter / \*

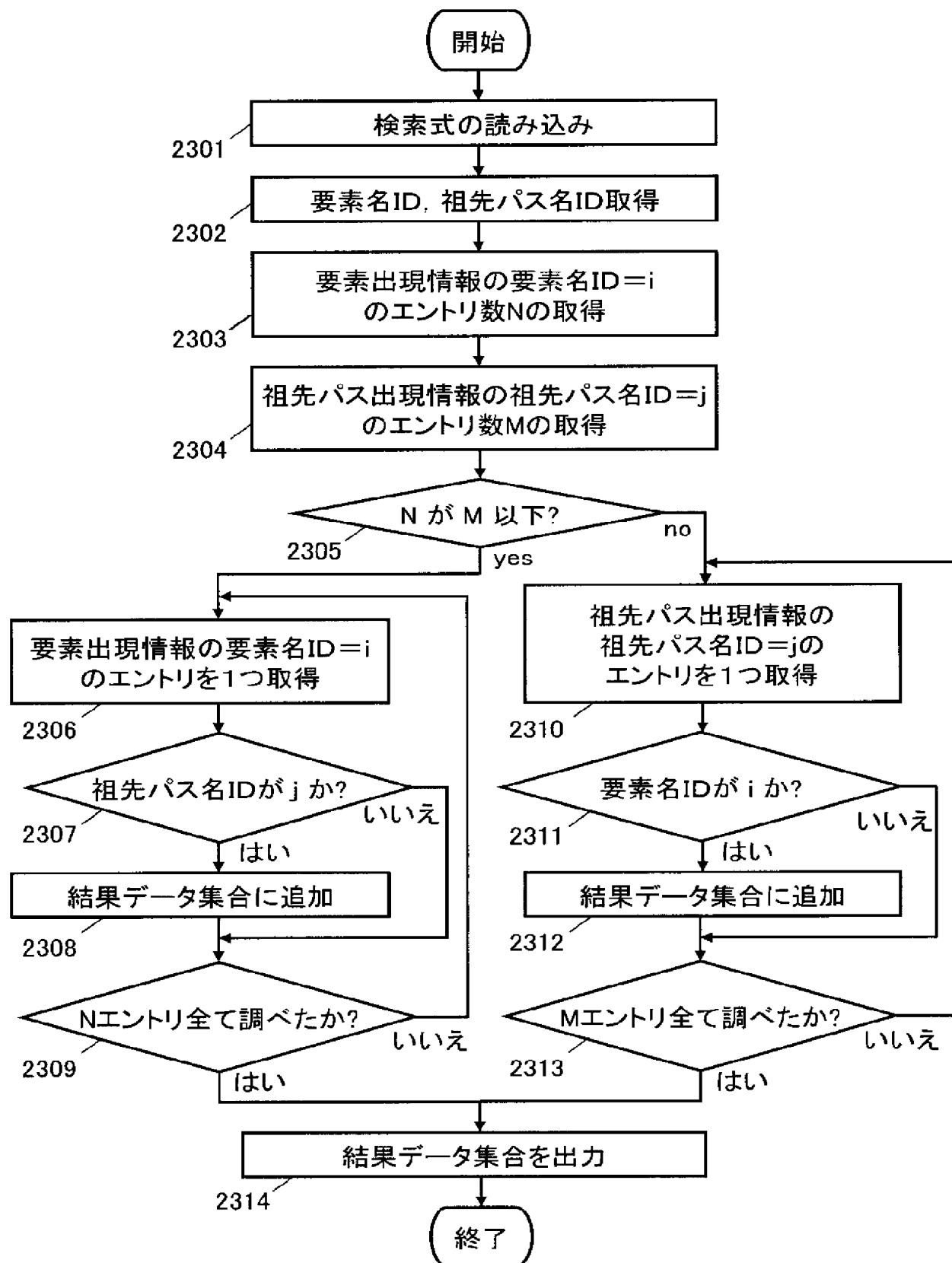
2103 // title

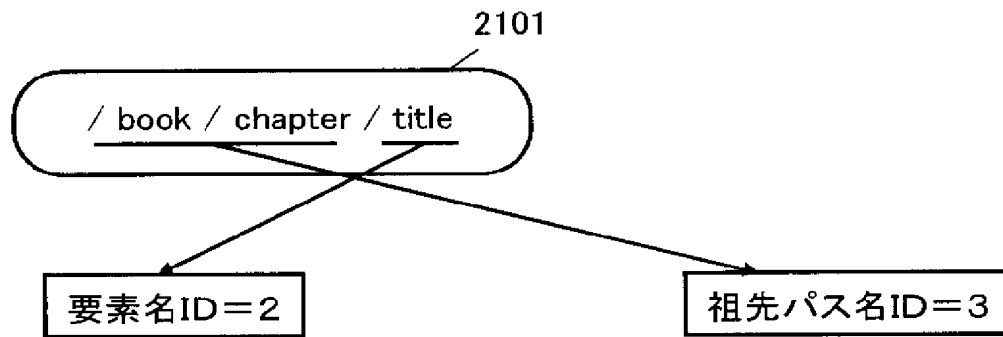
2104 / book / chapter / section[2]

2105 / book / chapter / section / @update

2106 / book / chapter / section [ contains( . , “極大単語”) ]

2107 / book / chapter / section / @update [ contains( . , “2004”) ]





## 要素出現情報

要素名ID	文書番号	文字位置	文字数	祖先パス名 ID	分岐順
2	1	0	4	2	1/1
	1	4	7	3	1/1/1
	1	59	4	3	1/2/1
	2	0	6	2	1/1
	4	0	8	2	1/1
	4	24	6	3	1/1/1
	7	0	5	2	1/1
	9	7	4	3	1/1/1
3					

1301

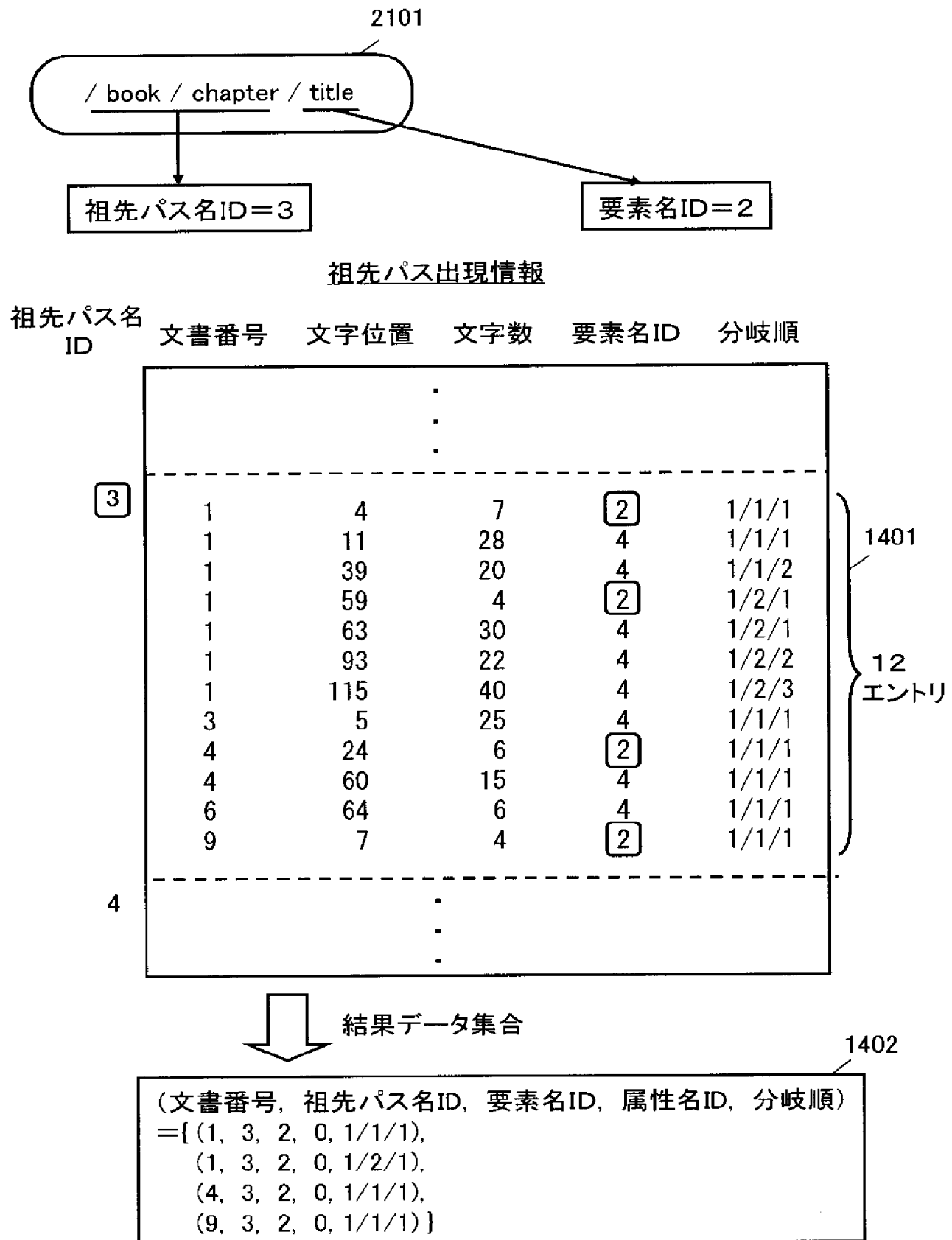
8  
エントリ



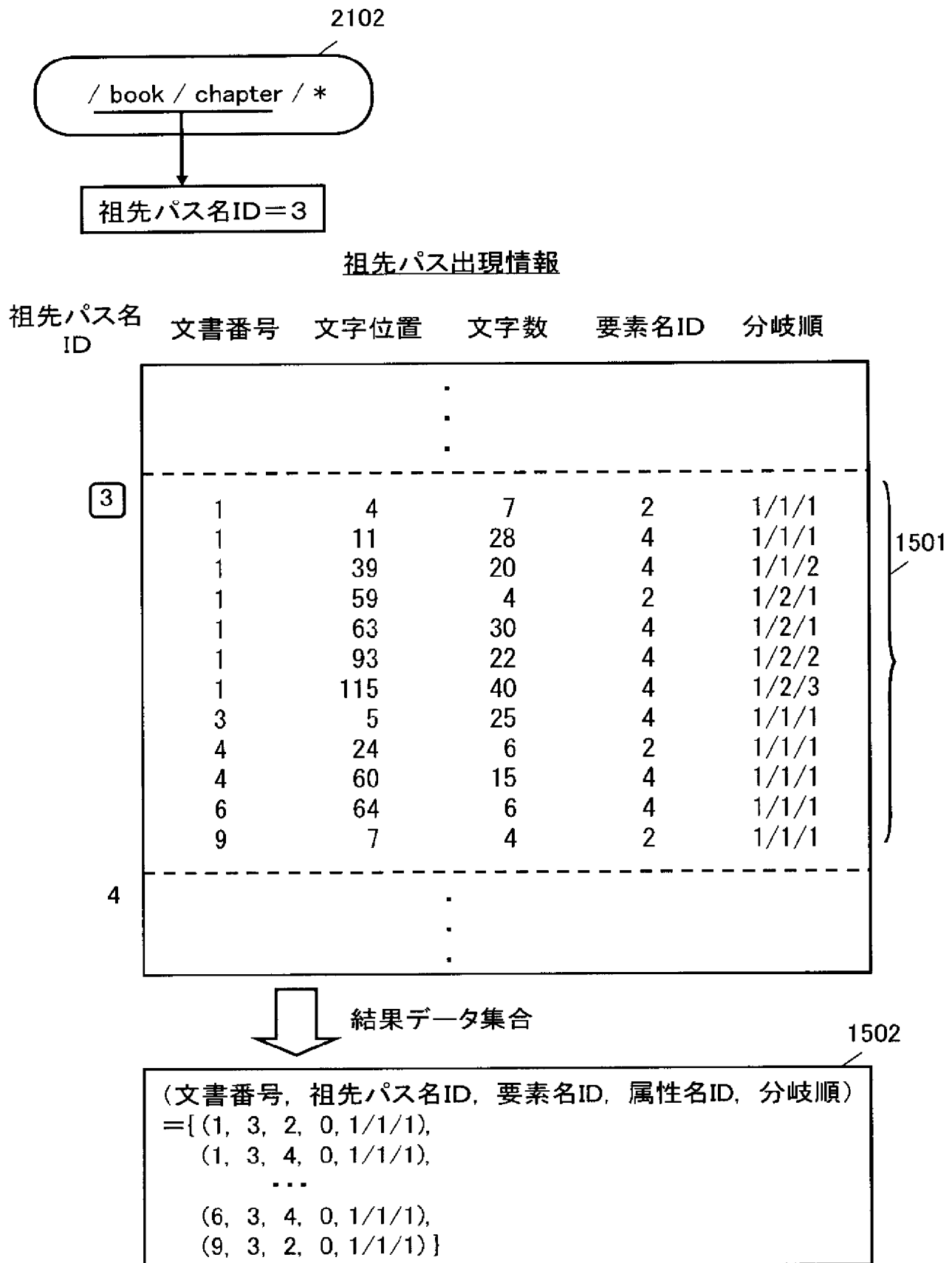
結果データ集合

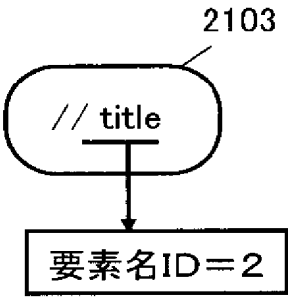
1302

(文書番号, 祖先パス名ID, 要素名ID, 属性名ID, 分岐順)  
 = { (1, 3, 2, 0, 1/1/1),  
 (1, 3, 2, 0, 1/2/1),  
 (4, 3, 2, 0, 1/1/1),  
 (9, 3, 2, 0, 1/1/1) }







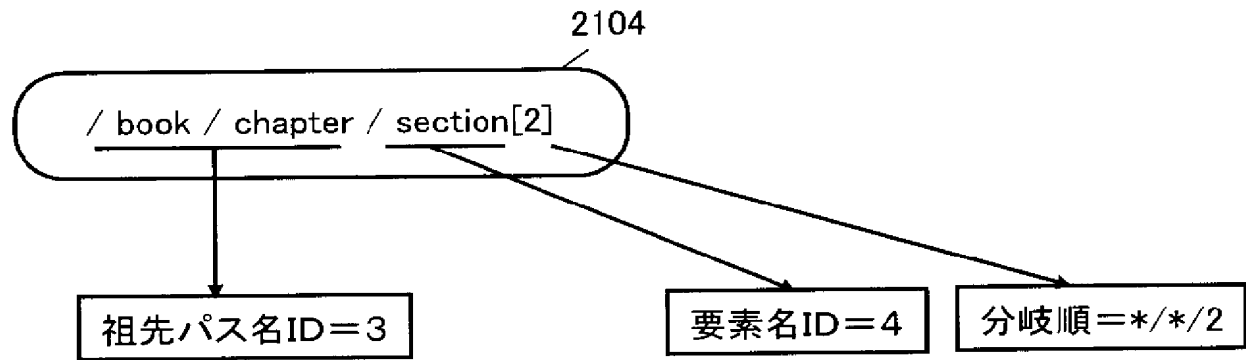


要素出現情報

要素名ID	文書番号	文字位置	文字数	祖先パス名 ID	分岐順
2	1	0	4	2	1/1
	1	4	7	3	1/1/1
	1	59	4	3	1/2/1
	2	0	6	2	1/1
	4	0	8	2	1/1
	4	24	6	3	1/1/1
	7	0	5	2	1/1
	9	7	4	3	1/1/1
3					

結果データ集合

(文書番号, 祖先パス名ID, 要素名ID, 属性名ID, 分岐順)  
=[ (1, 2, 2, 0, 1/1 ),  
  (1, 3, 2, 0, 1/1/1),  
  ...  
  (7, 2, 2, 0, 1/1 ),  
  (9, 3, 2, 0, 1/1/1) ]



## 祖先パス出現情報

祖先パス名ID 文書番号 文字位置 文字数 要素名ID 分岐順

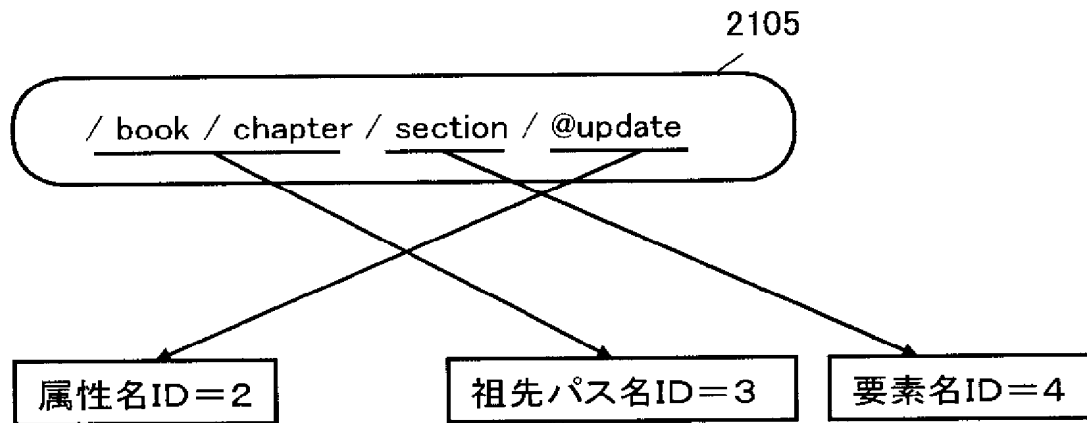
	1	.	.	.	.
	1	.	.	.	.
	1	.	.	.	.
3	1	4	7	2	1/1/1
	1	11	28	4	1/1/1
	1	39	20	4	1/1/2
	1	59	4	2	1/2/1
	1	63	30	4	1/2/1
	1	93	22	4	1/2/2
	1	115	40	4	1/2/3
	3	5	25	4	1/1/1
	4	24	6	2	1/1/1
	4	60	15	4	1/1/1
	6	64	6	4	1/1/1
	9	7	4	2	1/1/1
4		.	.	.	.
		.	.	.	.
		.	.	.	.

1701

結果データ集合

1702

(文書番号, 祖先パス名ID, 要素名ID, 属性名ID, 分岐順)  
 = { (1, 3, 4, 0, 1/1/2),  
 (1, 3, 4, 0, 1/2/2) }



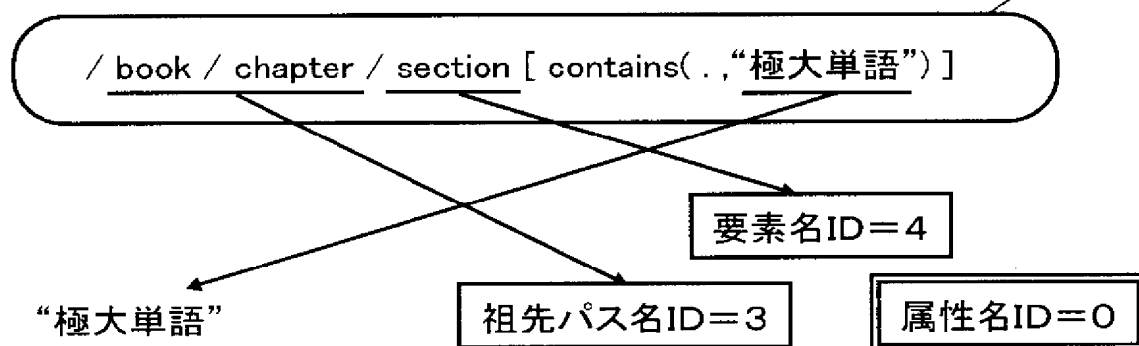
## 属性出現情報

属性名ID	文書番号	文字位置	文字数	祖先パス名 ID	要素名ID	分岐順	
							⋮
2	1	115	6	3	4	1/2/3	1801
	2	8	4	2	2	1/1	
	5	60	6	3	4	1/1/2	
	8	32	8	3	2	1/2/1	
3							⋮
							⋮
							⋮
							⋮

↓ 結果データ集合

1802

(文書番号, 祖先パス名ID, 要素名ID, 属性名ID, 分岐順)  
 = { (1, 3, 4, 2, 1/2/3),  
      (5, 3, 4, 2, 1/1/2) }



## テキスト出現情報

部分文字列 文書番号 文字位置 祖先パス名ID 要素名ID 属性名ID 分岐順

“極大”

1	118	3	4	0	1/2/3
2	86	3	4	0	1/1/1
3	24	2	2	0	1/1
4	62	3	4	0	1/1/1
8	77	3	4	2	1/1/1
		⋮			
		⋮			
		⋮			

文書番号が同じで文字位置が接続、分岐順も同じであること 1901

“単語”

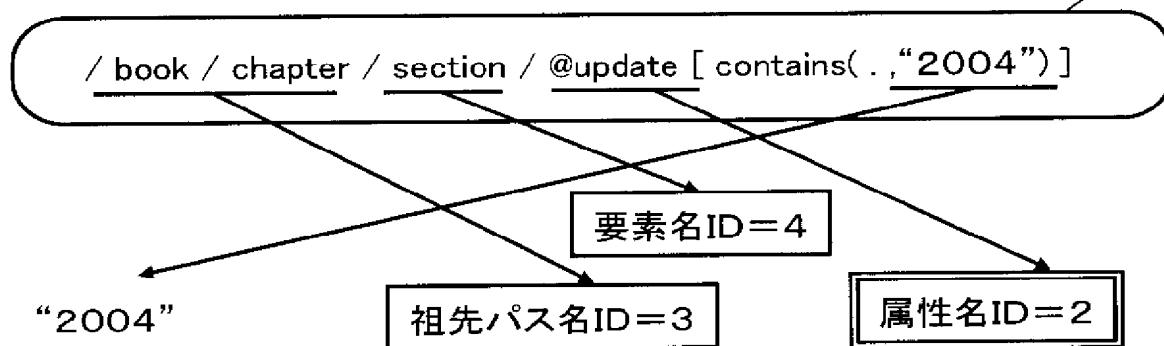
1	120	3	4	0	1/2/3
3	26	2	2	0	1/1
4	64	3	4	0	1/1/1
8	79	3	4	1	1/1/1
		⋮			
		⋮			
		⋮			

結果データ集合

1902

1903

(文書番号, 祖先パス名ID, 要素名ID, 属性名ID, 分岐順)  
 = { (1, 3, 4, 0, 1/2/3),  
 (4, 3, 4, 0, 1/1/1) }



## テキスト出現情報

部分文字列 文書番号 文字位置 祖先パス名 ID 要素名ID 属性名ID 分岐順

"20"

1	115	3	4	2	1/2/3
2	15	3	4	0	1/1/1
3	24	2	2	0	1/1
5	21	3	4	2	1/1/1
7	54	3	4	1	1/1/1
		⋮			
		⋮			
		⋮			

文書番号が同じで文字位置が接続、分岐順も同じであること 2001

"04"

1	117	3	4	2	1/2/3
3	26	2	2	0	1/1
5	23	3	4	2	1/1/1
7	56	3	4	1	1/1/1
		⋮			
		⋮			
		⋮			



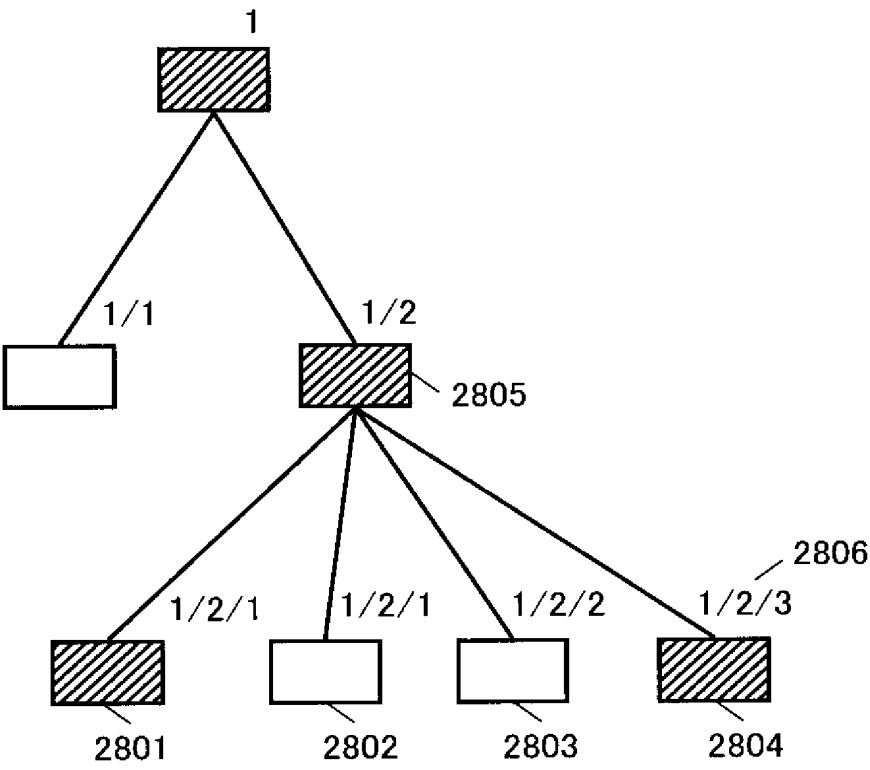
結果データ集合

2002

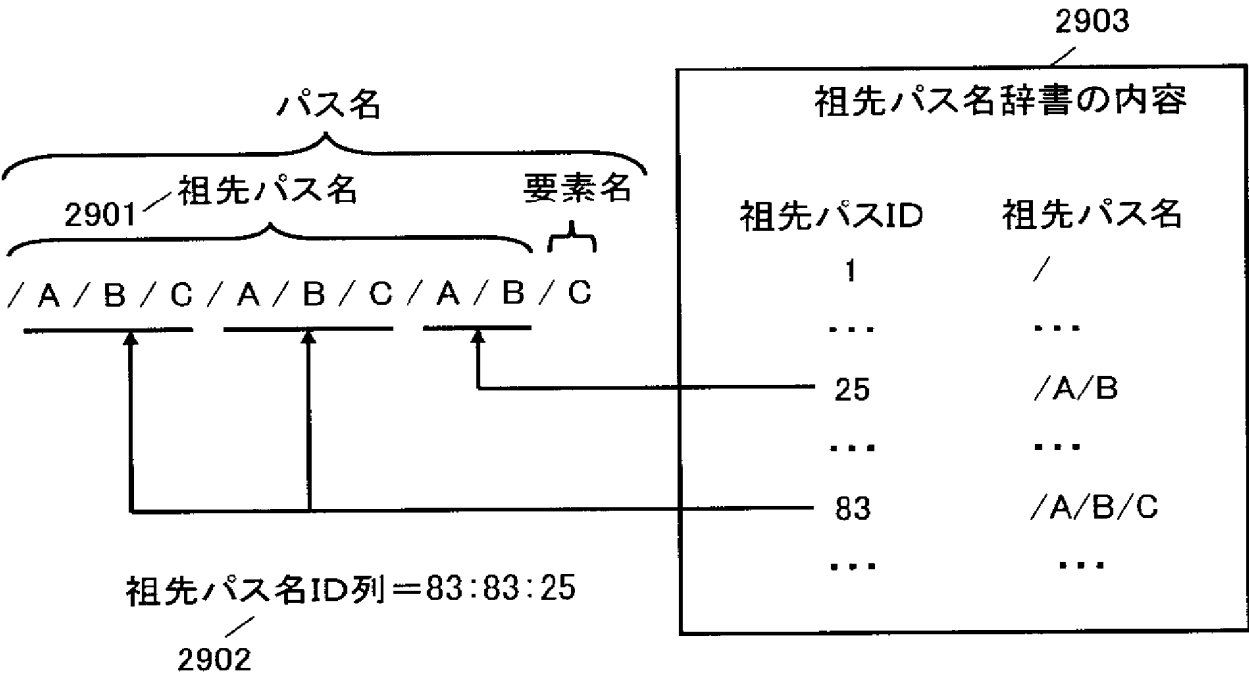
2003

(文書番号, 祖先パス名ID, 要素名ID, 属性名ID, 分岐順)  
 = { (1, 3, 4, 2, 1/2/3),  
 (5, 3, 4, 2, 1/1/1) }

【図 2 4】



【図 2 5】



【図 2 6】

要素出現情報

要素名ID	文書番号	文字位置	文字数	祖先パス名 ID	分岐順	空要素順
10	1	100	20	83:25	1/2/3/1/1/2	1/1/2/1/2/1

【図 2 7】

祖先パス出現情報

祖先パス名 ID	文書番号	文字位置	文字数	要素名ID	分岐順	空要素順
83:25	1	100	20	10	1/2/3/1/1/2	1/1/2/1/2/1

【図 2 8】

/ A / B / X / following-sibling::Y

3201

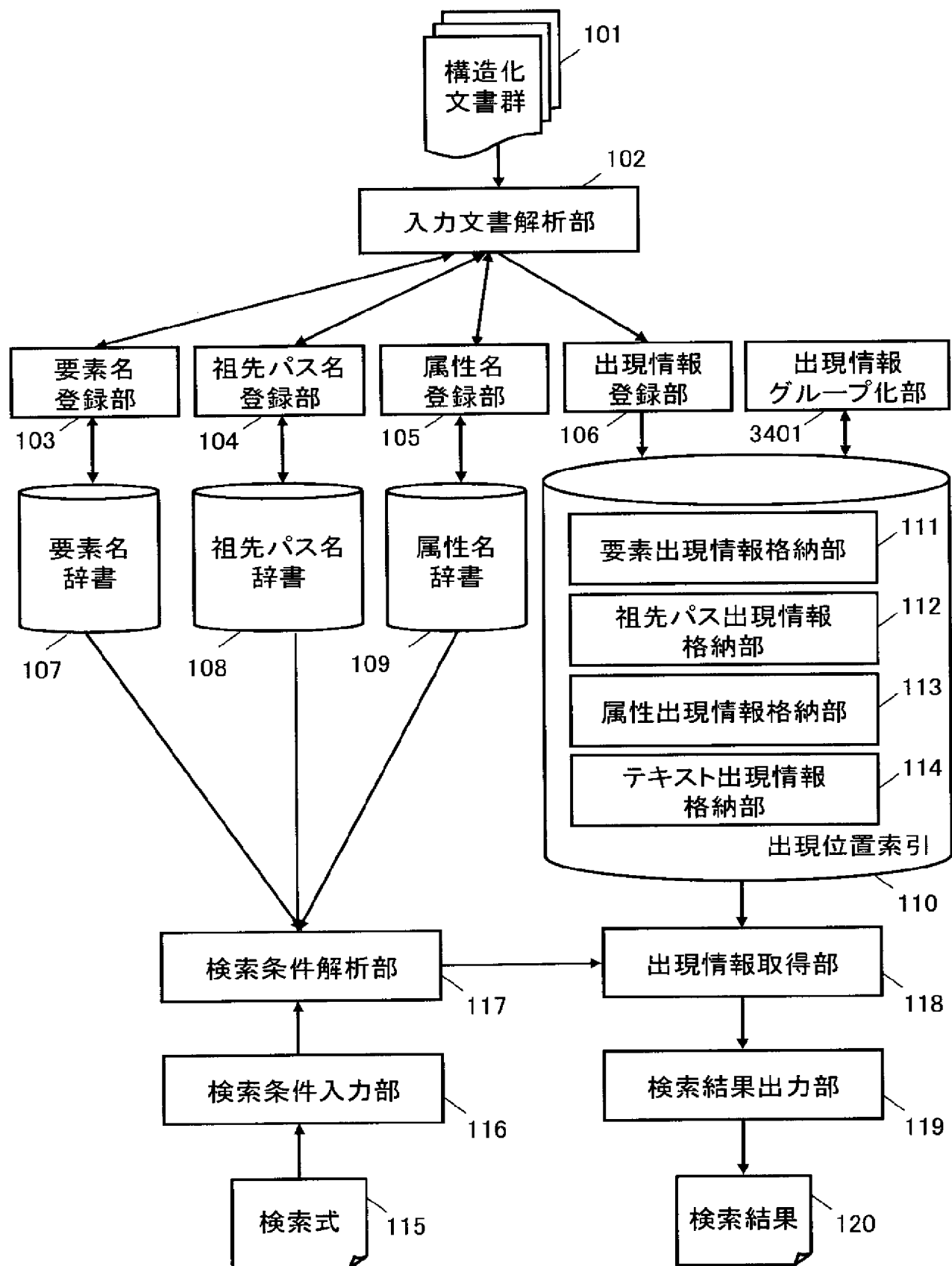


祖先パス出現情報

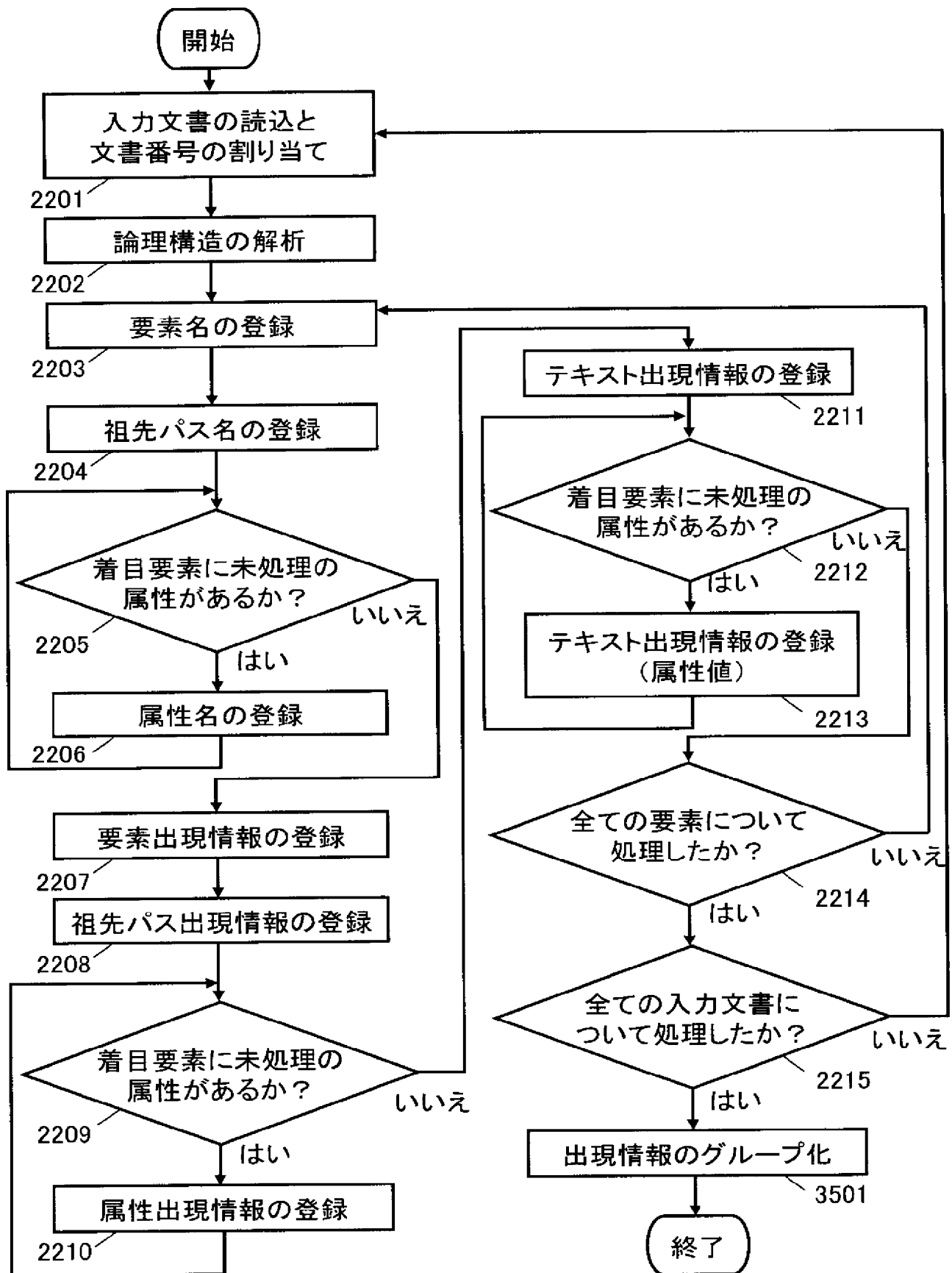
祖先パス名 ID	文書番号	文字位置	文字数	要素名ID	分岐順	空要素順
25 祖先パス名 /A/B	...					
	2	80	10	10 要素名 X	1/1/1	1/2/1
	...					
	2	120	20	14 要素名 Y	1/1/1	1/2/1
	...					
26	5	100	0	10	1/2/1	1/1/1
	5	100	30	14	1/2/1	1/1/2
	...					

結果データ集合

(文書番号, 祖先パス名ID, 要素名ID, 属性名ID, 分岐順, 空要素順)  
={ (2, 25, 14, 0, 1/1/1, 1/2/1),  
  (5, 25, 14, 0, 1/2/1, 1/1/2) }



【図 3 1】

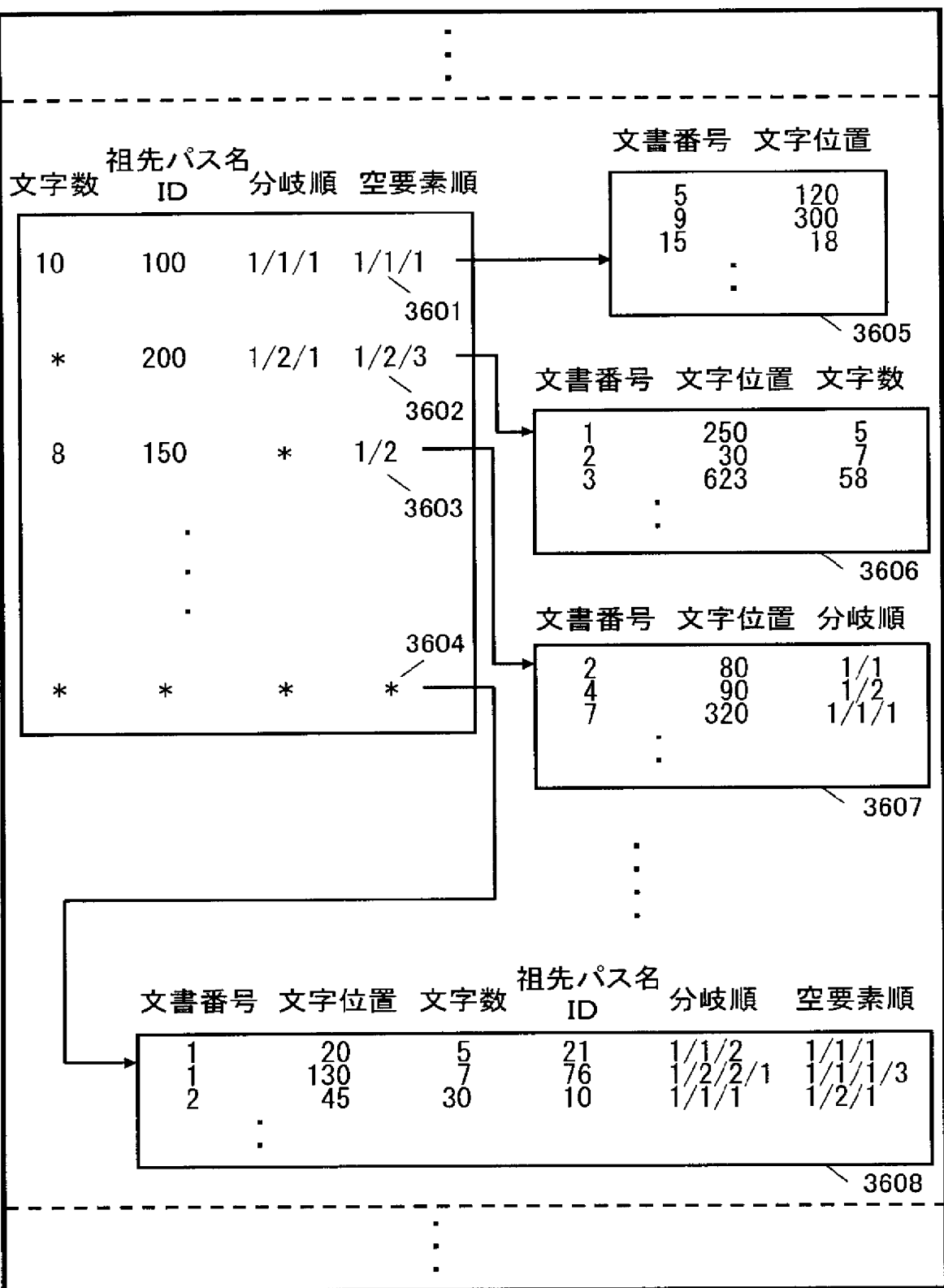


## グループ化された要素出現情報

要素名

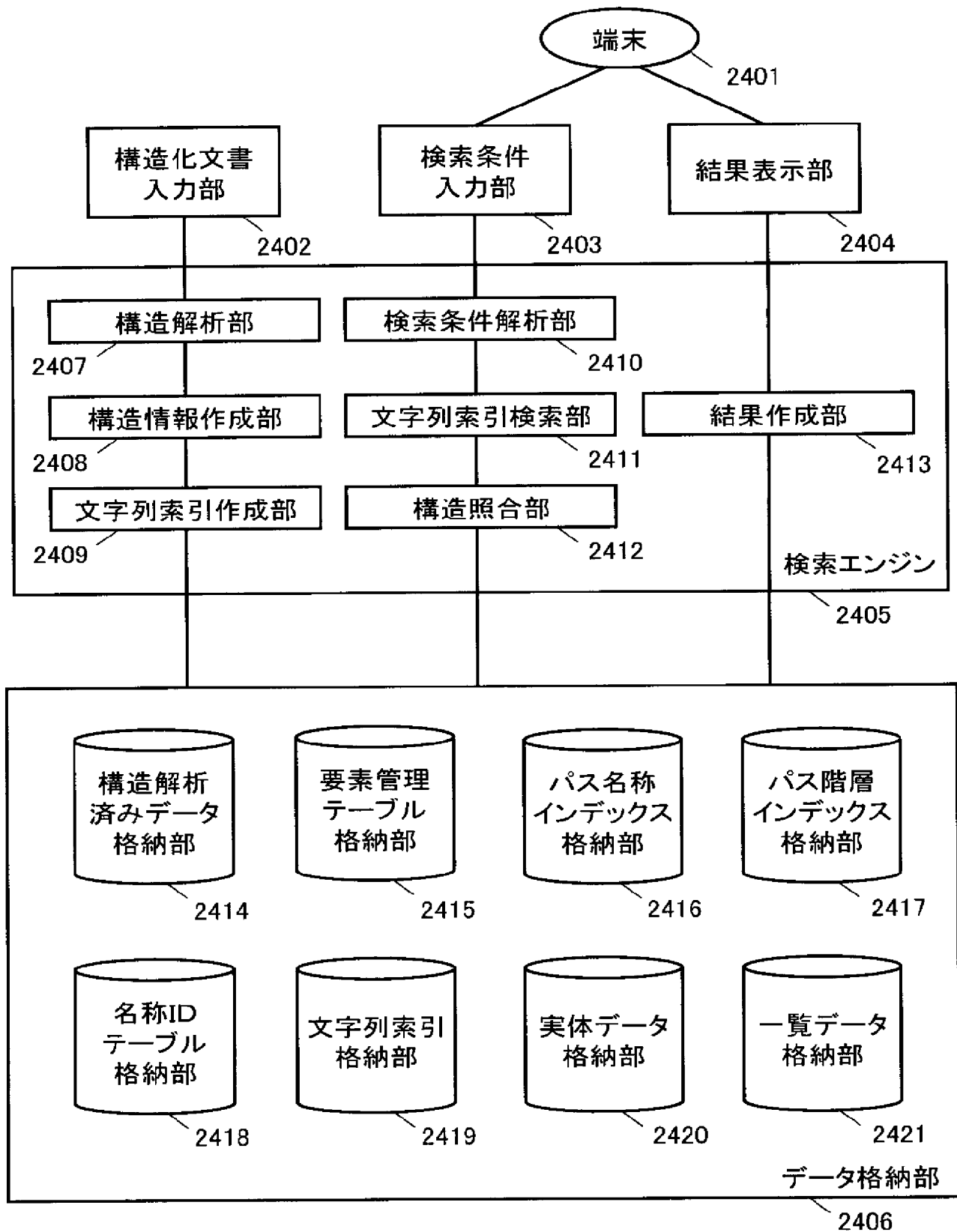
ID

14



15

【図 3 3】



検索単位  
識別子

文書番号

パス名称ID

パス階層ID

名称ID

1	1	N2	L2	T3
2	1	N3	L2	T4
3	1	N3	L6	T4
4	1	N4	L2	T5
5	1	N7	L3	T8
6	1	N8	L3	T9
7	1	N10	L4	T11
8	1	N11	L4	T9
・ ・ ・	...	...	...	...

要素管理テーブル

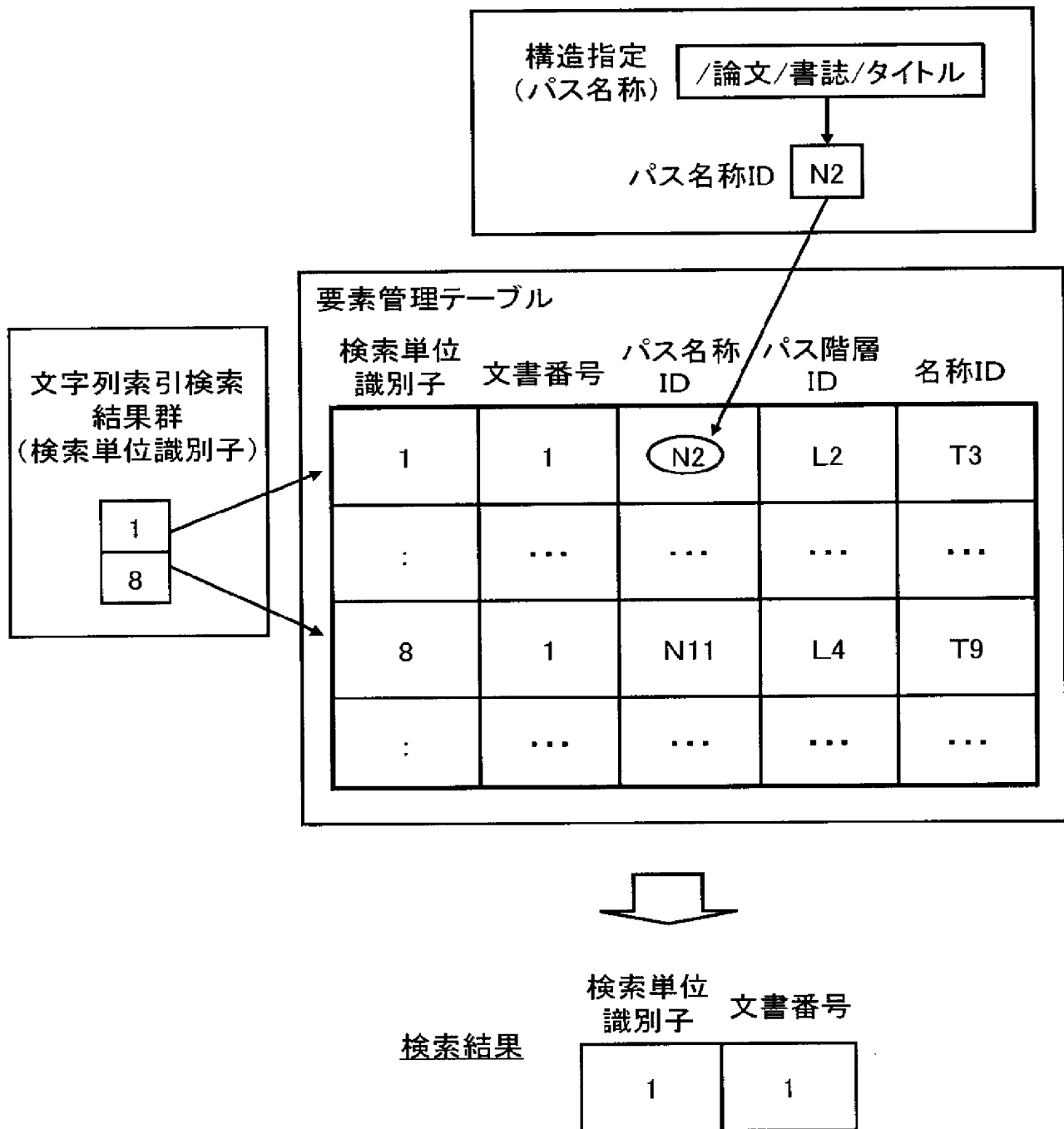
## &lt;タイトル&gt;構造化文書管理&lt;/タイトル&gt;

2601

“構造”	1	1
“造化”	1	2
“化文”	1	3
“文書”	1	4
“書管”	1	5
“管理”	1	6

検索単位  
識別子

文字位置番号





【書類名】 要約書

【要約】

【課題】 様々な検索条件での構造化文書に対する検索を効率良く行う。構造条件のみでの検索や、属性値に対する文字列検索ができるようにすることを目的とする。

【解決手段】 要素の出現情報を、要素名 I D をキーにして格納した要素出現情報格納部と、要素の出現情報を、その要素の祖先パス名 I D をキーにして格納した祖先パス出現情報格納部と、属性の出現情報を、属性名 I D をキーにして格納した属性出現情報格納部と、要素実体のテキスト文字列、および要素の持つ属性の属性値に関する出現情報を、部分文字列をキーにして格納したテキスト出現情報格納部とを備える。

【選択図】 図 1

## 出願人履歴

0 0 0 0 0 5 8 2 1

19900828

新規登録

大阪府門真市大字門真1 0 0 6 番地

松下電器産業株式会社